

68 5 81 04 84

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
2 December 2004 (02.12.2004)

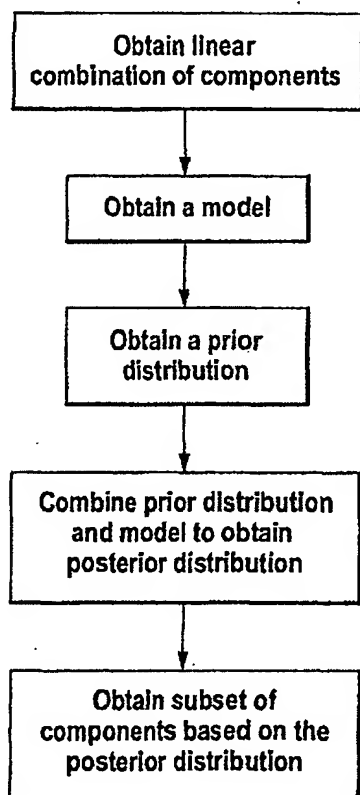
PCT

(10) International Publication Number
WO 2004/104856 A1

- (51) International Patent Classification⁷: **G06F 17/18**, 19/00 // 159:00
- (21) International Application Number: **PCT/AU2004/000696**
- (22) International Filing Date: **26 May 2004 (26.05.2004)**
- (25) Filing Language: **English**
- (26) Publication Language: **English**
- (30) Priority Data: **2003902589** **26 May 2003 (26.05.2003)** **AU**
- (71) Applicant (for all designated States except US): **COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANISATION [AU/AU]**; Limestone Avenue, Campbell, Australian Capital Territory 2612 (AU).
- (72) Inventors; and
(75) Inventors/Applicants (for US only): **KIIVERI, Harri [AU/AU]**; c/o- Commonwealth Scientific And Industrial Research Organisation, Limestone Avenue, Campbell, Australian Capital Territory 2612 (AU). **TRAJSTMAN, Albert [AU/AU]**; c/o- Commonwealth Scientific And Industrial Research Organisation, Limestone Avenue, Campbell, Australian Capital Territory 2612 (AU).
- (74) Agent: **GRIFFITH HACK**; GPO Box 4164, Sydney, New South Wales 2001 (AU).
- (81) Designated States (unless otherwise indicated, for every kind of national protection available): **AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM,**

[Continued on next page]

(54) Title: A METHOD FOR IDENTIFYING A SUBSET OF COMPONENTS OF A SYSTEM



(57) Abstract: A method of identifying a subset of components of a system based on data obtained from the system using at least one training sample from the system, the method comprising the steps of: obtaining a linear combination of components of the system and weightings of the linear combination of components, the weightings having values based on data obtained from the at least one training sample, the at least one training sample having a known feature; obtaining a model of a probability distribution of the known feature, wherein the model is conditional on the linear combination of components; obtaining a prior distribution for the weighting of the linear combination of the components, the prior distribution comprising a hyperprior having a high probability density close to zero, the hyperprior being such that it is not a Jeffreys hyperprior; combining the prior distribution and the model to generate a posterior distribution; and identifying the subset of components based on a set of the weightings that maximise the posterior distribution.

WO 2004/104856 A1



TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI,

Published:

— with international search report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

- 1 -

A METHOD FOR IDENTIFYING A SUBSET OF COMPONENTS OF A SYSTEM

FIELD OF THE INVENTION

5 The present invention relates to a method and apparatus for identifying components of a system from data generated from samples from the system, which components are capable of predicting a feature of the sample within the system and, particularly, but not exclusively, the present invention
10 relates to a method and apparatus for identifying components of a biological system from data generated by a biological method, which components are capable of predicting a feature of interest associated with a sample applied to the biological system.

15

BACKGROUND OF THE INVENTION

There are any number of systems in existence that can be classified according to one or more features thereof. The
20 term "system" as used throughout this specification is considered to include all types of systems from which data (e.g. statistical data) can be obtained. Examples of such systems include chemical systems, financial systems and geological systems. It is desirable to be able to utilise
25 data obtained from the systems to identify particular features of samples from the system; for instance, to assist with analysis of financial system to identify groups such as those who have good credit and those who are a credit risk. Often the data obtained from the systems is relatively large
30 and therefore it is desirable to identify components of the systems from the data, the components being predictive of the particular features of the samples from the system. However, when the data is relatively large it can be difficult to identify the components because there is a
35 large amount of data to process, the majority of which may not provide any indication or little indication of the features of a particular sample from which the data is

- 2 -

taken. Furthermore, components that are identified using a training sample are often ineffective at identifying features on test sample data when the test sample data has a high degree of variability relative to the training sample data. This is often the case in situations when, for example, data is obtained from many different sources, as it is often difficult to control the conditions under which the data is collected from each individual source.

10 An example of a type of system where these problems are particularly pertinent, is a biological system, in which the components could include, for example, particular genes or proteins. Recent advances in biotechnology have resulted in the development of biological methods for large scale
15 screening of systems and analysis of samples. Such methods include, for example, microarray analysis using DNA or RNA, proteomics analysis, proteomics electrophoresis gel analysis, and high throughput screening techniques. These types of methods often result in the generation of data that
20 can have up to 30,000 or more components for each sample that is tested.

It is highly desirable to be able identify features of interest in samples from biological systems. For example,
25 to classify groups such as "diseased" and "non-diseased". Many of these biological methods would be useful as diagnostic tools predicting features of a sample in the biological systems. For example, identifying diseases by screening tissues or body fluids, or as tools for
30 determining, for example, the efficacy of pharmaceutical compounds.

Use of biological methods such as biotechnology arrays in such applications to date has been limited due to the large
35 amount of data that is generated from these types of methods, and the lack of efficient methods for screening the data for meaningful results. Consequently, analysis of

- 3 -

biological data using existing methods is time consuming,
prone to false results and requires large amounts of
computer memory if a meaningful result is to be obtained
from the data. This is problematic in large scale screening
5 scenarios where rapid and accurate screening is required.

It is therefore desirable to have a method, in particular
for analysis of biological data, and more generally, for an
improved method of analysing data from a system in order to
10 predict a feature of interest for a sample from the system.

SUMMARY OF THE INVENTION

According to a first aspect of the present invention, there
15 is provided a method of identifying a subset of components
of a system based on data obtained from the system using at
least one training sample from the system, the method
comprising the steps of:

obtaining a linear combination of components of the
20 system and weightings of the linear combination of
components, the weightings having values based on the data
obtained from the system using the at least one training
sample, the at least one training sample having a known
feature;

25 obtaining a model of a probability distribution of the
known feature, wherein the model is conditional on the
linear combination of components;

obtaining a prior distribution for the weighting of
the linear combination of the components, the prior
30 distribution comprising a hyperprior having a high
probability density close to zero, the hyperprior being such
that it is not a Jeffreys hyperprior;

combining the prior distribution and the model to
generate a posterior distribution; and

35 identifying the subset of components based on a set of
the weightings that maximise the posterior distribution.

distribution having two possible classes or groups such as an on/off state. Examples of such groups include dead/alive, improved/not improved, depressed/not depressed. A multinomial distribution is a generalisation of the

5 binomial distribution in which a plurality of classes or groups are possible for each of a plurality of samples, or in other words, a sample may be classified into one of a plurality of classes or groups. Thus, by defining a likelihood function based on a multinomial or binomial

10 logistic regression, it is possible to identify subsets of components that are capable of classifying a sample into one of a plurality of pre-defined groups or classes. To do this, training samples are grouped into a plurality of sample groups (or "classes") based on a predetermined

15 feature of the training samples in which the members of each sample group have a common feature and are assigned a common group identifier. A likelihood function is formulated based on a multinomial or binomial logistic regression conditional on the linear combination (which incorporates the data

20 generated from the grouped training samples). The feature may be any desired classification by which the training samples are to be grouped. For example, the features for classifying tissue samples may be that the tissue is normal, malignant, benign, a leukemia cell, a healthy cell, that the

25 training samples are obtained from the blood of patients having or not having a certain condition, or that the training samples are from a cell from one of several types of cancer as compared to a normal cell.

30 In the first embodiment, the likelihood function based on the multinomial or binomial logistic regression is of the form:

$$L = \prod_{i=1}^n \left(\prod_{g=1}^{G-1} \left\{ \frac{e^{x_i^T \beta_g}}{1 + \sum_{g=1}^{G-1} e^{x_i^T \beta_g}} \right\}^{e_{ik}} \left\{ \frac{1}{1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}} \right\}^{e_{iG}} \right)$$

γ_{ik} is the probability that training sample i belongs to a
35 class with identifier less than or equal to k (where the

- 8 -

total of ordered classes is G). The r_i is defined further in the document.

In a third embodiment of the present invention, the
5 likelihood function is based on the generalised linear model. The generalised linear model preferably models a feature that is distributed as a regular exponential family of distributions. Examples of regular exponential family of distributions include normal distribution, gaussian
10 distribution, poisson distribution, gamma distribution and inverse gaussian distribution. Thus, in another embodiment of the method of the invention, a subset of components is identified that is capable of predicting a predefined characteristic of a sample which has a distribution
15 belonging to a regular exponential family of distributions. In particular by defining a generalised linear model which models the characteristic to be predicted. Examples of a characteristic that may be predicted using a generalised linear model include any quantity of a sample that exhibits
20 the specified distribution such as, for example, the weight, size or other dimensions or quantities of a sample.

In the third embodiment, the generalised linear model is of the form:

$$L = \log p(y | \beta, \phi) = \sum_{i=1}^N \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

25

where $y = (y_1, \dots, y_n)^T$ and $a_i(\phi) = \phi / w_i$ with the w_i being a fixed set of known weights and ϕ a single scale parameter. The other terms in this expression are defined later in this document.

30

In a fourth embodiment, the method of the present invention may be used to predict the time to an event for a sample by utilising the likelihood function that is based on a hazard model, which preferably estimates the probability of a time
35 to an event given that the event has not taken place at the

time of speeding up the data. In the fourth embodiment, the likelihood function is selected from the group comprising a Cox's proportional hazards model, parametric survival model and accelerated failure times model. Cox's proportional hazards model permits the time to an event to be modelled on a set of components and component weights without making restrictive assumptions about time. The accelerated failure model is a general model for data consisting of survival times in which the component measurements are assumed to act multiplicatively on the time-scale, and so affect the rate at which an individual proceeds along the time axis. Thus, the accelerated survival model can be interpreted in terms of the speed of progression of, for example, disease. The parametric survival model is one in which the distribution function for the time to an event (eg survival time) is modelled by a known distribution or has a specified parametric formulation. Among the commonly used survival distributions are the Weibull, exponential and extreme value distributions.

20 In the fourth embodiment, a subset of components capable of predicting the time to an event for a sample is identified by defining a likelihood based on Cox's proportional standards model, a parametric survival model or an
25 accelerated survival times model, which comprises measuring the time elapsed for a plurality of samples from the time the sample is obtained to the time of the event.

In the fourth embodiment, the likelihood function for
30 predicting the time to an event is of the form:

$$\text{Log (Partial) Likelihood} = \sum_{i=1}^N g_i(\underline{\beta}, \underline{\varphi}; X, y, \underline{c})$$

where $\underline{\beta}' = (\beta_1, \beta_2, \dots, \beta_p)$ and $\underline{\varphi}' = (\varphi_1, \varphi_2, \dots, \varphi_q)$ are the model parameters, y is a vector of observed times and c is an indicator vector which indicates whether a time is a true

For any defined models, the weightings are typically estimated using a Bayesian statistical model (Kotz and Johnson, 1983) in which a posterior distribution of the component weights is formulated which combines the likelihood function and a prior distribution. The component weightings are estimated by maximising the posterior

30 where $\beta^T = (\beta_1^T, \dots, \beta_{G-1}^T)$ and $\tau^T = (\tau_1^T, \dots, \tau_{G-1}^T)$.

- 12 -

and $p(\beta_g | \tau_g^2)$ is $N(0, \text{diag}\{\tau_g^2\})$ and $P(\tau_g^2)$ is some hyperprior distribution for τ_g^2 .

As used herein, when the likelihood function for the probability distribution is based on a categorical ordered logistic regression, the notation for the prior distribution is:

$$P(\beta_1, \beta_2, \dots, \beta_n) = \int \prod_{i=1}^N P(\beta_i | v_i^2) P(v_i^2) dv^2$$

where $\beta_1, \beta_2, \dots, \beta_n$ are component weights, $P(\beta_i | v_i)$ is $N(0, v_i^2)$ and $P(v_i)$ some hyperprior distribution for v_i .

As used herein, when the likelihood function for the distribution is based on a generalised linear model, the notation for the prior distribution is:

$$p(\beta) = \int p(\beta | v^2) p(v^2) dv^2$$

wherein v is a $p \times 1$ vector of hyperparameters, and where $p(\beta | v^2)$ is $N(0, \text{diag}\{v^2\})$ and $p(v^2)$ is some prior distribution for v^2 .

As used herein, when the likelihood function for the distribution is based on a hazard model, the notation for the prior distribution is:

where $p(\beta^* | \tau)$ is $N(0, \text{diag}\{\tau^2\})$ and $p(\tau)$ some hyperprior distribution for τ .

During the iterative procedure, component weightings having a value less than a pre-determined threshold are eliminated, preferably by setting those component weights to zero. This results in the substantially elimination of the

- 5 The EM algorithm produces a sequence of component weighting estimates that converge to give component the weightings that maximise the probability density of the posterior distribution. The EM algorithm consists of two steps, known as the E or Expectation step and the M, or Maximisation
10 step. In the E step, the expected value of the log-posterior function conditional on the observed data is determined. In the M step, the expected log-posterior function is maximised to give updated component weight estimates that increase the posterior. The two steps are
15 alternated until convergence of the E step and the M step is achieved, or in other words, until the expected value and the maximised value of the expected log-posterior function converges.
- 20 It is envisaged that the method of the present invention may be applied to any system from which measurements can be obtained, and preferably systems from which very large amounts of data are generated. Examples of systems to which the method of the present invention may be applied include
25 biological systems, chemical systems, agricultural systems, weather systems, financial systems including, for example, credit risk assessment systems, insurance systems, marketing systems or company record systems, electronic systems, physical systems, astrophysics systems and mechanical
30 systems. For example, in a financial system, the samples may be particular stock and the components may be measurements made on any number of factors which may affect stock prices such as company profits, employee numbers, rainfall values in various cities, number of shareholders
35 etc.

- 16 -

protein electrophoresis gels and proteomics electrophoresis gels.

The components may be any measurable component of the
 5 system. In the case of a biological system, the components
 may be, for example, genes or portions thereof, DNA
 sequences, RNA sequences, peptides, proteins, carbohydrate
 molecules, lipids or mixtures thereof, physiological
 components, anatomical components, epidemiological
 10 components or chemical components.

The training samples may be any data obtained from a system
 in which the feature of the sample is known. For example,
 training samples may be data generated from a sample applied
 15 to a biological system. For example, when the biological
 system is a DNA microarray, the training sample may be data
 obtained from the array following hybridisation of the array
 with RNA extracted from cells having a known feature, or
 cDNA synthesised from the RNA extracted from cells, or if
 20 the biological system is a proteomics electrophoresis gel,
 the training sample may be generated from a protein or cell
 extract applied to the system.

It is envisaged that an embodiment of a method of the
 25 present invention may be used in re-evaluating or evaluating
 test data from subjects who have presented mixed results in
 response to a test treatment. Thus, there is a second
 aspect to the present invention.

30 The second aspect provides a method for identifying a subset
 of components of a subject which are capable of classifying
 the subject into one of a plurality of predefined groups,
 wherein each group is defined by a response to a test
 treatment, the method comprising the steps of:

35 exposing a plurality of subjects to the test treatment
 and grouping the subjects into response groups based on
 responses to the treatment;

WO 2004/104856

PCT/AU2004/000696

- 17 -

measuring components of the subjects; and
identifying a subset of components that is capable of
classifying the subjects into response groups using a
statistical analysis method.

5

Preferably, the statistical analysis method comprises the method according to the first aspect of the present invention.

10 Once a subset of components has been identified, that subset
can be used to classify subjects into groups such as those
that are likely to respond to the test treatment and those
that are not. In this manner, the method of the present
invention permits treatments to be identified which may be
15 effective for a fraction of the population, and permits
identification of that fraction of the population that will
be responsive to the test treatment.

According to a third aspect of the present invention, there is provided an apparatus for identifying a subset of components of a subject, the subset being capable of being used to classify the subject into one of a plurality of predefined response groups wherein each response group, is formed by exposing a plurality of subjects to a test treatment and grouping the subjects into response groups based on the response to the treatment, the apparatus comprising:

an input for receiving measured components of the subjects; and

30 processing means operable to identify a subset of
 components that is capable of being used to classify the
 subjects into response groups using a statistical analysis
 method.

35 Preferably, the statistical analysis method comprises the method according to the first or second aspect.

- 19 -

carbohydrates, lipids or any other measurable component of the subject.

In a particularly embodiment of the fifth aspect, the
5 compound is a pharmaceutical compound or a composition comprising a pharmaceutical compound and a pharmaceutically acceptable carrier.

The identification method of the present invention may be
10 implemented by appropriate computer software and hardware.

According to a sixth aspect of the present invention, there is provided an apparatus for identifying a subset of components of a system from data generated from the system
15 from a plurality of samples from the system, the subset being capable of being used to predict a feature of a test sample, the apparatus comprising:

a processing means operable to:

obtain a linear combination of components of the system
20 and obtain weightings of the linear combination of components, each of the weightings having a value based on data obtained from at least one training sample, the at least one training sample having a known feature;

obtaining a model of a probability distribution of a
25 second feature, wherein the model is conditional on the linear combination of components;

obtaining a prior distribution for the weightings of the linear combination of the components, the prior distribution comprising an adjustable hyperprior which
30 allows the prior probability mass close to zero to be varied wherein the hyperprior is not a Jeffrey's hyperprior;

combining the prior distribution and the model to generate a posterior distribution; and

identifying the subset of components having component
35 weights that maximize the posterior distribution.

- 20 -

Preferably, the processing means comprises a computer arranged to execute software.

According to a seventh aspect of the present invention,
5 there is provided a computer program which, when executed by a computing apparatus, allows the computing apparatus to carry out the method according to the first aspect of the present invention.

10 The computer program may implement any of the preferred algorithms and method steps of the first or second aspect of the present invention which are discussed above.

According to an eighth aspect of the present invention,
15 there is provided a computer readable medium comprising the computer program according with the seventh aspect of the present invention.

According to a ninth aspect of the present invention, there
20 is provided a method of testing a sample from a system to identify a feature of the sample, the method comprising the steps of testing for a subset of components that are diagnostic of the feature, the subset of components having been determined by using the method according to the first
25 or second aspect of the present invention.

Preferably, the system is a biological system.

According to a tenth aspect of the present invention, there
30 is provided an apparatus for testing a sample from a system to determine a feature of the sample, the apparatus comprising means for testing for components identified in accordance with the method of the first or second aspect of the present invention.

35

According to an eleventh aspect of the present invention, there is provided a computer program which, when executed by

- 21 -

on a computing device, allows the computing device to carry out a method of identifying components from a system that are capable of being used to predict a feature of a test sample from the system, and wherein a linear combination of components and component weights is generated from data generated from a plurality of training samples, each training sample having a known feature, and a posterior distribution is generated by combining a prior distribution for the component weights comprising an adjustable hyperprior which allows the probability mass close to zero to be varied wherein the hyperprior is not a Jeffrey's hyperprior, and a model that is conditional on the linear combination, to estimate component weights which maximise the posterior distribution.

Where aspects of the present invention are implemented by way of a computing device, it will be appreciated that any appropriate computer hardware e.g. a PC or a mainframe or a networked computing infrastructure, may be used.

According to a twelfth aspect of the present invention, there is provided a method of identifying a subset of components of a biological system, the subset being capable of predicting a feature of a test sample from the biological system, the method comprising the steps of:

obtaining a linear combination of components of the system and weightings of the linear combination of components, each of the weightings having a value based on data obtained from at least one training sample, the at least one training sample having a known first feature;

obtaining a model of a probability distribution of a second feature, wherein the model is conditional on the linear combination of components;

obtaining a prior distribution for the weightings of the linear combination of the components, the prior distribution comprising an adjustable hyperprior which allows the probability mass close to zero to be varied;

- 22 -

combining the prior distribution and the model to generate a posterior distribution; and

identifying the subset of components based on the weightings that maximize the posterior distribution.

5

BRIEF DESCRIPTION OF THE DRAWINGS

Notwithstanding any other embodiments that may fall within the scope of the present invention, an embodiment of the present invention will now be described, by way of example only, with reference to the accompanying figures, in which:

10

figure 1 provides a flow chart of a method according to the embodiment of the present invention;

15

figure 2 provides a flow chart of another method according to the embodiment of the present invention;

figure 3 provides a block diagram of an apparatus according to the embodiment of the present invention;

20

figure 4 provides a flow chart of a further method according to the embodiment of the present invention;

figure 5 provides a flow chart of an additional method according to the embodiment of the present invention; and

25

figure 6 provides a flow chart of yet another method according to the embodiment of the present invention.

30

DETAILED DESCRIPTION OF AN EMBODIMENT

The embodiment of the present invention identifies a relatively small number of components which can be used to identify whether a particular training sample has a feature. The components are "diagnostic" of that feature, or enable discrimination between samples having a different feature.

35

- 23 -

The number of components selected by the method can be controlled by the choice of parameters in the hyperprior. It is noted that the hyperprior is a gamma distribution with a specified shape and scale parameter. Essentially, from all the data which is generated from the system, the method of the present invention enables identification of a relatively small number of components which can be used to test for a particular feature. Once those components have been identified by this method, the components can be used in future to assess new samples. The method of the present invention utilises a statistical method to eliminate components that are not required to correctly predict the feature.

The inventors have found that component weightings of a linear combination of components of data generated from the training samples can be estimated in such a way as to eliminate the components that are not required to correctly predict the feature of the training sample. The result is that a subset of components are identified which can correctly predict the feature of the training sample. The method of the present invention thus permits identification from a large amount of data a relatively small and controllable number of components which are capable of correctly predicting a feature.

The method of the present invention also has the advantage that it requires usage of less computer memory than prior art methods. Accordingly, the method of the present invention can be performed rapidly on computers such as, for example, laptop machines. By using less memory, the method of the present invention also allows the method to be performed more quickly than other methods which use joint (rather than marginal) information on components for analysis of, for example, biological data.

The samples are grouped into sample groups (or "classes") based on a pre-determined classification. The classification

- 25 -

may be any desired classification by which the training samples are to be grouped. For example, the classification may be whether the training samples are from a leukemia cell or a healthy cell, or that the training samples are obtained
 5 from the blood of patients having or not having a certain condition, or that the training samples are from a cell from one of several types of cancer as compared to a normal cell.

In one embodiment, the input data is organised into an
 10 $n \times p$ data matrix $X = (x_{ij})$ with n training samples and p components. Typically, p will be much greater than n .

In another embodiment, data matrix X may be replaced by an $n \times n$ kernel matrix K to obtain smooth functions of X as
 15 predictors instead of linear predictors. An example of the kernel matrix K is $k_{ij} = \exp(-0.5 * (x_i - x_j)^t (x_i - x_j) / \sigma^2)$ where the subscript on x refers to a row number in the matrix X . Ideally, subsets of the columns of K are selected which give sparse representations of these smooth functions.

20 Associated with each sample class (group) may be a class label y_i , where $y_i = k, k \in \{1, \dots, G\}$, which indicates which of G sample classes a training sample belongs to. We write the $n \times 1$ vector with elements y_i as \underline{y} . Given the vector \underline{y} we
 25 can define indicator variables

$$e_{ig} = \begin{cases} 1, & y_i = g \\ 0, & \text{otherwise} \end{cases} \quad (A1)$$

In one embodiment, the component weights are estimated using a Bayesian statistical model (see Kotz and Johnson, 1983).
 30 Preferably, the weights are estimated by maximising the posterior distribution of the weights given the data generated from each training sample. This results in an objective function to be maximised consisting of two parts. The first part a likelihood function and the second a prior
 35 distribution for the weights which ensures that zero weights are preferred whenever possible. In a preferred embodiment,

the likelihood function is derived from a multiclass logistic model. Preferably, the likelihood function is computed from the probabilities:

$$p_{ig} = \frac{e^{x_i^T \beta_g}}{\left(1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}\right)}, g = 1, \dots, G-1 \quad (A2)$$

5 and

$$p_{iG} = \frac{1}{\left(1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}\right)} \quad (A3)$$

wherein

- 10 p_{ig} is the probability that the training sample with input data X_i will be in sample class g ;
 $x_i^T \beta_g$ is a linear combination generated from input data from training sample i with component weights β_g ;
 x_i^T is the components for the i^{th} Row of X and β_g is a set of
 15 component weights for sample class g ;

Typically, as discussed above, the component weights are estimated in a manner which takes into account the apriori assumption that most of the component weights are zero.

20

In one embodiment, components weights β_g in equation (A2) are estimated in a manner whereby most of the values are zero, yet the samples can still be accurately classified.

- 25 In one embodiment, the component weights are estimated by maximising the posterior distribution of the weights given the data in the Bayesian model referred to above.

Preferably, the component weights are estimated by

30

PCT/AU2004/000696

- 27 -

- 5

10

(A4)

15

In one embodiment, $p(\tau_g^2) = \prod_{i=1}^n p(\tau_{ig}^2)$ where $p(\tau_{ig}^2)$ is a prior

wherein $t_{ig}^2 = 1/\tau_{ig}^2$ has an independent gamma distribution.

20

25

(A5)

30

In one embodiment, the first derivative is determined from the following algorithm:

$$\frac{\partial \log L}{\partial \beta_g} = X^T (\underline{e}_g - p_g), \quad g=1, \dots, G-1 \quad (A6)$$

wherein $\underline{e}_g^T = (e_{ig}, i=1, \dots, n)$, $p_g^T = (p_{ig}, i=1, \dots, n)$ are vectors indicating
 5 membership of sample class g and probability of class g respectively.

In one embodiment, the second derivative is determined from the following algorithm:

10

$$\frac{\partial^2 \log L}{\partial \beta_g \partial \beta_h} = -X^T \text{diag} \{ \delta_{hg} p_g - p_h p_g \} X \quad (A7)$$

where δ_{hg} is 1 if h equals g and zero otherwise.

15 Equation A6 and equation A7 may be derived as follows:

(a) Using equations (A1), (A2) and (A3), the likelihood function of the data can be written as:

20

$$L = \prod_{i=1}^n \left(\prod_{g=1}^{G-1} \left\{ \frac{e^{x_i^T \beta_g}}{1 + \sum_{g=1}^{G-1} e^{x_i^T \beta_g}} \right\}^{e_{ig}} \left\{ \frac{1}{1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}} \right\}^{e_{iG}} \right) \quad (A8)$$

(b) Taking logs of equation (A6) and using the fact that

$\sum_{h=1}^G e_{ih} = 1$ for all i gives:

25

$$\log L = \sum_{i=1}^n \left(\sum_{g=1}^{G-1} e_{ig} x_i^T \beta_g - \log \left(1 + \sum_{g=1}^{G-1} e^{x_i^T \beta_g} \right) \right) \quad (A9)$$

(c) Differentiating equation (A8) with respect to β_g gives

- 29 -

$$\frac{\partial \log L}{\partial \beta_g} = X^T (e_g - p_g), \quad g = 1, \dots, G-1 \quad (A10)$$

whereby $e_g^T = (e_{ig}, i=1, n)$, $p_g^T = (p_{ig}, i=1, n)$ are vectors indicating membership of sample class g and probability of class g respectively.

(d) The second derivative of equation (9) has elements

$$\frac{\partial^2 \log L}{\partial \beta_g \partial \beta_h} = -X^T \text{diag} \{ \delta_{hg} p_g - p_h p_g \} X \quad (A11)$$

where

$$\delta_{hg} = \begin{cases} 1, & h = g \\ 0, & \text{otherwise} \end{cases}$$

Component weights which maximise the posterior distribution of the likelihood function may be specified using an EM algorithm comprising an E step and an M step.

In conducting the EM algorithm, the E step preferably comprises the step computing a term of the form:

$$\begin{aligned} P &= \sum_{g=1}^{G-1} \sum_{i=1}^n E \{ \beta_{ig}^2 / \tau_{ig}^2 \mid \hat{\beta}_{ig} \} \\ &= \sum_{g=1}^{G-1} \sum_{i=1}^n \beta_{ig}^2 / \hat{d}_{ig}^2 \end{aligned} \quad (A11a)$$

where $\hat{d}_{ig} = E \{ 1 / \tau_{ig}^2 \mid \hat{\beta}_{ig} \}^{-0.5}$, $\hat{d}_g = (\hat{d}_{1g}, \hat{d}_{2g}, \dots, \hat{d}_{pg})^T$ and $\hat{d}_{ig} = 1 / \hat{d}_{ig} = 0$ if $\hat{\beta}_{ig} = 0$.

Preferably, equation (11a) is computed by calculating the conditional expected value of $t_{ig}^2 = 1 / \tau_{ig}^2$ when $p(\beta_{ig} \mid \tau_{ig}^2)$ is $N(0, \tau_{ig}^2)$ and $p(\tau_{ig}^2)$ has a specified prior distribution.

Explicit formulae for the conditional expectation will be presented later.

5 Typically, the EM algorithm comprises the steps:

- (a) performing an E step by calculating the conditional expected value of the posterior distribution of component weights using the function:

$$Q=Q(\gamma|y,\hat{\gamma})=\log L-\frac{1}{2}\sum_{g=1}^{G-1}\gamma_g^T\text{diag}\{d_g(\hat{\gamma}_g)\}^{-2}\gamma_g \quad (\text{A12})$$

where $x_i^T\beta_g=x_i^TP_g\gamma_g$ in equation (8), $d(\hat{\gamma}_g)=P_g^T\hat{d}_g$, and \hat{d}_g is defined as in equation (11a) evaluated at $\hat{\beta}_g=P_g\hat{\gamma}_g$. Here P_g is a matrix of zeroes and ones derived from the identity matrix such that $P_g^T\beta_g$ selects non-zero elements of β_g which are denoted by γ_g .

- (b) performing an M step by applying an iterative procedure to maximise Q as a function of γ whereby:

$$\gamma^{t+1}=\gamma^t-\alpha'\left(\frac{\partial^2 Q}{\partial \gamma^2}\right)^{-1}\left(\frac{\partial Q}{\partial \gamma}\right) \quad (\text{A13})$$

where α' is a step length such that $0\leq\alpha'\leq 1$; and $\gamma=(\gamma_g, g=1,\dots,G-1)$.

30 Equation (A12) may be derived as follows:

Calculate the conditional expected value of (A5) given the observed data y and a set of parameter estimates $\hat{\beta}$.

$$= \begin{bmatrix} X_1^T (e_1 - p_1) \\ \vdots \\ X_{G-1}^T (e_{G-1} - p_{G-1}) \end{bmatrix} - \text{diag} \{ d(\hat{\gamma}) \}^{-2} \gamma \quad (\text{A15})$$

and

$$\frac{\partial^2 Q}{\partial \gamma^2} = \left(\frac{\partial \beta}{\partial \gamma} \right) \frac{\partial^2 \log L}{\partial \beta^2} \left(\frac{\partial \beta}{\partial \gamma} \right)^T - \text{diag} \{d(\hat{\gamma})\}^{-2}$$

5

$$= - \left\{ \begin{pmatrix} X_1^T \Delta_{11} X_1 & \dots & X_1^T \Delta_{1G-1} X_{G-1} \\ \vdots & & \vdots \\ X_{G-1} \Delta_{G-11} X_1 & & X_{G-1} \Delta_{G-1G-1} X_{G-1} \end{pmatrix} + \text{diag} \{d(\hat{\gamma})\}^{-2} \right\} \quad (\text{A16})$$

where

$$\begin{aligned} \Delta_{gh} &= \text{diag} \{ \delta_{gh} p_g - p_g p_h \}, \\ \delta_{gh} &= \begin{cases} 1, & g = h \\ 0, & \text{otherwise} \end{cases} \\ d(\hat{\gamma}) &= (d(\hat{\gamma}_g), g = 1, \dots, G-1) \end{aligned}$$

10 and

$$X_g^T = P_g^T X^T, g = 1, \dots, G-1 \quad (\text{A17})$$

In a preferred embodiment, the iterative procedure may be
15 simplified by using only the block diagonals of equation
(A16) in equation (A13). For $g=1, \dots, G-1$, this gives:

$$\gamma_g^{t+1} = \gamma_g' + \alpha' \left\{ X_g^T \Delta_{gg} X_g + \text{diag} \{d_g(\hat{\gamma}_g)\}^{-2} \right\}^{-1} \left\{ X_g^T (e_g - p_g) - \text{diag} \{d_g(\hat{\gamma}_g)\} \gamma_g' \right\} - \quad (\text{A18})$$

20

Rearranging equation (A18) leads to

$$\gamma_g^{t+1} = \gamma_g' + \alpha' \text{diag} \{d_g(\hat{\gamma}_g)\} \left(Y_g^T \Delta_{gg} Y_g + I \right)^{-1} \left\{ Y_g^T (e_g - p_g) - \text{diag} \{d_g(\hat{\gamma}_g)\}^{-1} \gamma_g' \right\}^{-1} \quad (\text{A19})$$

25 where

$$Y_g^T = \text{diag} \{d_g(\hat{\gamma}_g)\} X_g^T$$

PCT/AU2004/000696

- 33 -

5 noting that:

(A20)

10 (A19) with (A20) substituted into equation (A19) is used

Note that when τ_{ig}^2 has a Jeffreys prior we have:

15 $E\{t_{ig}^2 | \hat{\beta}_{ig}\} = 1/\hat{\beta}_{ig}^2$

$k > 0$ so that the density of t_{ig}^2 is:

$$\gamma(t_{ig}^2, b, k) = b^{-1} (t_{ig}^2/b)^{k-1} \exp(-t_{ig}^2/b) / \Gamma(k)$$

shown that

$$E\{t^2 | \beta\} = (2k+1)/(2/b + \beta^2) \quad (A21)$$

as follows:

30 Define

$$I(p,b,k)=\int_0^{\infty}(t^2)^p t \exp(-0.5\beta^2 t^2)\gamma(t^2,b,k)dt^2$$

then

$$I(p,b,k)=b^{p+0.5} \{ \Gamma(p+k+0.5)/\Gamma(k) \} (1+0.5b\beta^2)^{-(p+k+0.5)}$$

Proof

5 Let $s = \beta^2 / 2$ then

$$I(p,b,k)=b^{p+0.5} \int_0^{\infty} (t^2/b)^{p+0.5} \exp(-st^2) \gamma(t^2,b,k) dt^2$$

Now using the substitution $u = t^2/b$ we get

$$I(p,b,k)=b^{p+0.5} \int_0^{\infty} (u)^{p+0.5} \exp(-sub) \gamma(u,1,k) du$$

10 Now let $s'=bs$ and substitute the expression for $\gamma(u,1,k)$. This gives

$$I(p,b,k)=b^{p+0.5} \int_0^{\infty} \exp(-(s'+1)u) u^{p+k+0.5-1} du / \Gamma(k)$$

Looking up a table of Laplace transforms, eg Abramowitz and Stegun, then gives the result.

15 The conditional expectation follows from

$$\begin{aligned} E\{t^2 | \beta\} &= I(1,b,k)/I(0,b,k) \\ &= (2k+1)/(2/b + \beta^2) \end{aligned}$$

20 As k tends to zero and b tends to infinity we get the equivalent result using Jeffreys prior. For example, for $k=0.005$ and $b=2*10^5$

$$E\{t^2 | \beta\} = (1.01)/(10^{-5} + \beta^2)$$

25 Hence we can get arbitrarily close to the Jeffreys prior with this proper prior.

The algorithm for this model has

WO 2004/104856

PCT/AU2004/000696

- 35 -

$$\begin{aligned}\hat{d} &= E\{t^2 | \hat{\beta}\}^{-0.5} \\ &= E\left\{\frac{1}{\tau^2} | \hat{\beta}\right\}^{-0.5}\end{aligned}$$

where the expectation is calculated as above.

In another embodiment, τ_{ig}^2 has an independent gamma distribution with scale parameter $b > 0$ and shape parameter $k > 0$. It can be shown that

$$\begin{aligned}E\{\tau_{ig}^{-2} | \beta_{ig}\} &= \frac{\int_0^\infty u^{k-3/2-1} \exp(-(\gamma_{ig}/u + u)) du}{b \int_0^\infty u^{k-1/2-1} \exp(-(\gamma_{ig}/u + u)) du} \\ &= \sqrt{\frac{2}{b}} \frac{1}{|\beta_{ig}|} \frac{K_{3/2-k}(2\sqrt{\gamma_{ig}})}{K_{1/2-k}(2\sqrt{\gamma_{ig}})} \\ &= \frac{1}{|\beta_{ig}|^2} \frac{(2\sqrt{\gamma_{ig}}) K_{3/2-k}(2\sqrt{\gamma_{ig}})}{K_{1/2-k}(2\sqrt{\gamma_{ig}})}\end{aligned}\tag{A22}$$

where $\gamma_{ig} = \beta_{ig}^2/2b$ and K denotes a modified Bessel function. For $k=1$ in equation (A22)

$$E\{\tau_{ig}^{-2} | \beta\} = \sqrt{2/b} (1/|\beta_{ig}|)$$

For $K=0.5$ in equation (A22)

$$E\{\tau_{ig}^{-2} | \beta_{ig}\} = \sqrt{2/b} (1/|\beta_{ig}|) \{K_1(2\sqrt{\gamma_{ig}})/K_0(2\sqrt{\gamma_{ig}})\}$$

or equivalently

$$E\{\tau_{ig}^{-2} | \beta_i\} = (1/|\beta_{ig}|^2) \{2\sqrt{\gamma_i} K_1(2\sqrt{\gamma_{ig}})/K_0(2\sqrt{\gamma_{ig}})\}$$

Proof of (A.1)

From the definition of the conditional expectation, writing

$\gamma = \beta^2/2b$, we get

- 36 -

$$E\{\tau^{-2}|\beta\} = \frac{\int_0^{\infty} \tau^{-2} \tau^{-1} \exp(-\gamma \tau^{-2}) b^{-1} (\tau^{-2}/b)^{k-1} \exp(\tau^{-2}/b) d\tau^2}{\int_0^{\infty} \tau^{-1} \exp(-\gamma \tau^{-2}) b^{-1} (\tau^{-2}/b)^{k-1} \exp(\tau^{-2}/b) d\tau^2}$$

Rearranging, simplifying and making the substitution $u=\tau^2/b$ gives the first equation in (A22).

5 The integrals in (22) can be evaluated by using the result

$$\int_0^{\infty} x^{-b-1} \exp\left[-\left(x + \frac{a^2}{x}\right)\right] dx = \frac{2}{a^b} K_b(2a)$$

where K denotes a modified Bessel function, see Watson(1966).

10

Examples of members of this class are $k=1$ in which case

$$E\{\tau_{ig}^{-2}|\beta_{ig}\} = \sqrt{2/b}(1/|\beta_{ig}|)$$

15 which corresponds to the prior used in the Lasso technique, Tibshirani(1996). See also Figueiredo(2001).

The case $k=0.5$ gives

$$20 \quad E\{\tau_{ig}^{-2}|\beta_{ig}\} = \sqrt{2/b}(1/|\beta_{ig}|) \{K_1(2\sqrt{\gamma_{ig}})/K_0(2\sqrt{\gamma_{ig}})\}$$

or equivalently

$$E\{\tau_{ig}^{-2}|\beta_{ig}\} = (1/|\beta_{ig}|^2) \{2\sqrt{\gamma_{ig}} K_1(2\sqrt{\gamma_{ig}})/K_0(2\sqrt{\gamma_{ig}})\}$$

25

where K_0 and K_1 are modified Bessel functions, see Abramowitz and Stegun(1970). Polynomial approximations for evaluating these Bessel functions can be found in Abramowitz and Stegun(1970, p379). The expressions above demonstrate the connection with the Lasso model and the Jeffreys prior model.

30

35

classified as normal, benign, localised tumour or metastasised tumour. Thus, the subset of components is diagnostic of whether a test sample belongs to a particular class within an ordered set of classes. It will be apparent
5 that once the subset of components have been identified, only the subset of components need be tested in future diagnostic procedures to determine to what ordered class a sample belongs.

10 The method of the invention is particularly suited for the analysis of very large amounts of data. Typically, large data sets obtained from test samples is highly variable and often differs significantly from that obtained from the training samples. The method of the present invention is
15 able to identify subsets of components from a very large amount of data generated from training samples, and the subset of components identified by the method can then be used to classifying test samples even when the data generated from the test sample is highly variable compared
20 to the data generated from training samples belonging to the same class. Thus, the method of the invention is able to identify a subset of components that are more likely to classify a sample correctly even when the data is of poor quality and/or there is high variability between samples of
25 the same ordered class.

The components are "predictive" for that particular ordered class. Essentially, from all the data which is generated from the system, the method of the present invention enables
30 identification of a relatively small number of components which can be used to classify the training data. Once those components have been identified by this method, the components can be used in future to classify test samples. The method of the present invention preferably utilises a
35 statistical method to eliminate components that are not required to correctly classify the sample into a class that is a member of an ordered class.

- 40 -

In the following there are N samples, and vectors such as y , z and μ have components y_i , z_i and μ_i for $i = 1, \dots, N$. Vector multiplication and division is defined componentwise and
 5 $\text{diag}\{ \cdot \}$ denotes a diagonal matrix whose diagonals are equal to the argument. We also use $\| \cdot \|$ to denote Euclidean norm.

Preferably, there are N observations y_i , where y_i takes
 10 integer values $1, \dots, G$. The values denote classes which are ordered in some way such as for example severity of disease. Associated with each observation there is a set of covariates (variables, e.g. gene expression values) arranged into a matrix X^* with n row and p columns wherein n is the
 15 samples and p the components. The notation x_i^T denotes the i^{th} row of X^* . Individual (sample) i has probabilities of belonging to class k given by $\pi_{ik} = \pi_i(x_i^T)$. Define cumulative probabilities

$$\gamma_{ik} = \sum_{g=1}^k \pi_{ig}, \quad k = 1, \dots, G$$

20 Note that γ_{ik} is just the probability that observation i belongs to a class with index less than or equal to k . Let C be a N by p matrix with elements c_{ij} given by

$$c_{ij} = \begin{cases} 1, & \text{if observation } i \text{ in class } j \\ 0, & \text{otherwise} \end{cases}$$

25 and let R be an n by P matrix with elements r_{ij} given by

$$r_{ij} = \sum_{g=1}^j c_{ig}$$

These are the cumulative sums of the columns of C within
 rows.

30

For independent observations (samples) the likelihood of the data may be written as:

Following Figueiredo(2001), in order to eliminate redundant variables (covariates), a prior is specified for the

- 42 -

parameters β^* by introducing a $p \times 1$ vector of hyperparameters.

Preferably, the prior specified for the component weights is
5 of the form

$$p(\beta^*) = \int_{\tau^2} p(\beta^* | \tau^2) p(\tau^2) d\tau^2 \quad (B1)$$

where $p(\beta^* | \tau^2)$ is $N(0, \text{diag}\{\tau^2\})$ and $p(\tau^2)$ is a suitably chosen hyperprior. For example, $p(\tau^2) \propto \prod_{i=1}^n p(\tau_i^2)$ is a suitable form of
10 Jeffreys prior.

In another embodiment, $p(\tau_i^2)$ is a prior wherein $\tau_i^2 = 1/\nu_i^2$ has an independent gamma distribution.

15 In another embodiment, $p(\nu_i^2)$ is a prior wherein ν_i^2 has an independent gamma distribution.

The elements of theta have a non informative prior.

20 Writing $L(y | \beta^* \theta)$ for the likelihood function, in a Bayesian framework the posterior distribution of β , θ and ν given y is

$$p(\beta^* \theta \nu | y) \propto L(y | \beta^* \theta) p(\beta^* | \nu) p(\nu) \quad (2)$$

25

By treating ν as a vector of missing data, an iterative algorithm such as an EM algorithm (Dempster et al, 1977) can be used to maximise (2) to produce maximum a posteriori estimates of β and θ . The prior above is such that the

where the ridge parameter λ satisfies $0 < \lambda \leq 1$ and ζ is small and chosen so that the link function $g(z) = \log(z/(1-z))$ is well defined at $z = y + \zeta$.

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

and let P_n be a matrix of zeroes and ones such that the nonzero elements $\gamma^{(n)}$ of $\beta^{(n)}$ satisfy

$$\begin{aligned} \gamma^{(n)} &= P_n^T \beta^{(n)}, & \beta^{(n)} &= P_n \gamma^{(n)} \\ \gamma &= P_n^T \beta, & \beta &= P_n \gamma \end{aligned}$$

Define $w_\beta = (w_{\beta i}, i=1, p)$, such that

$$w_{\beta i} = \begin{cases} 1, & i \geq G \\ 0, & \text{otherwise} \end{cases}$$

and let $w_\gamma = P_n w_\beta$

3. Perform the E step by calculating

$$\begin{aligned} Q(\beta | \beta^{(n)}, \phi^{(n)}) &= E\{\log p(\beta, \phi, v | y) | y, \beta^{(n)}, \phi^{(n)}\} \\ &= L(y | \beta, \phi^{(n)}) - 0.5 (\|(\beta^* w_\beta) / \hat{d}^{(n)}\|^2) \end{aligned} \quad (15)$$

where L is the log likelihood function of y and

$$\hat{d}_{ig} = E\{1/\tau_{ig}^2 | \hat{\beta}_{ig}\}^{-0.5}, \hat{d}_g = (\hat{d}_{1g}, \hat{d}_{2g}, \dots, \hat{d}_{pg})^T \text{ and for convenience we}$$

define $\hat{d}_{ig} = 1/\hat{d}_{ig} = 0$ if $\hat{\beta}_{ig} = 0$. Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ (15)

can be written as

$$Q(\gamma | \gamma^{(n)}, \phi^{(n)}) = L(y | P_n \gamma, \phi^{(n)}) - 0.5 (\|(\gamma^* w_\gamma) / d(\gamma^{(n)})\|^2) \quad (B4)$$

with $d(\gamma^{(n)}) = P_n^T d^{(n)}$ evaluated at $\beta^{(n)} = P_n \gamma^{(n)}$.

4. Do the M step. This can be done with Newton Raphson iterations as follows. Set $\gamma_0 = \gamma^{(n)}$ and for $r=0, 1, 2, \dots$ γ_{r+1}

WO 2004/104856

PCT/AU2004/000696

- 45 -

= $\gamma_r + \alpha_r \delta_r$ where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1} | \gamma^{(n)}, \phi^{(n)}) > Q(\gamma_r | \gamma^{(n)}, \phi^{(n)})$.

For $p \leq N$ use

$$\delta_r = \Delta(d^*(\gamma^{(n)})) [Y_n^T V_r^{-1} Y_n + I]^{-1} (Y_n^T z_r - \frac{w_r \gamma_r}{d^*(\gamma^{(n)})}) \quad (B5)$$

5 where

$$d^*(\gamma^{(n)}) = \begin{cases} d(\gamma^{(n)}), & i \geq G \\ \kappa, & \text{otherwise} \end{cases}$$

$$Y_n^T = \Delta(d^*(\gamma^{(n)})) P_n^T X^T$$

10 $V_r^{-1} = \text{diag}\{\mu_r, (1-\mu_r)\}$

$$z_r = (y - \mu_r)$$

and $\mu_r = \exp(X P_n \gamma_r) / (1 + \exp(X P_n \gamma_r))$.

15

For $p > N$ use

$$\delta_r = \Delta(d^*(\gamma^{(n)})) [I - Y_n^T (Y_n Y_n^T + V_r)^{-1} Y_n] (Y_n^T z_r - \frac{w_r \gamma_r}{d^*(\gamma^{(n)})}) \quad (B6)$$

with V_r and z_r defined as before.

20

Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g

$$|| \gamma_r - \gamma_{r+1} || < \epsilon \quad (\text{for example } 10^{-5}).$$

25 5. Define $\beta^* = P_n \gamma^*$, $S_{n+1} = \{i \geq G: |\beta_i| > \max_{j \geq G} |\beta_j| * \epsilon_1\}$ where ϵ_1 is a small constant, say $1e-5$. Set $n=n+1$.

6. Check convergence. If $|| \gamma^* - \gamma^{(n)} || < \epsilon_2$ where ϵ_2 is suitably small then stop, else go to step 2 above.

30 Recovering the probabilities

Once we have obtained estimates of the parameters β are obtained, calculate

$$a_{ik} = \frac{\hat{\pi}_{ik}}{\hat{\gamma}_{ik}}$$

for $i=1, \dots, N$ and $k = 2, \dots, G$.

Preferably, to obtain the probabilities we use the recursion

$$\pi_{iG} = a_{iG}$$

$$\pi_{ik-1} = \left(\frac{a_{ik-1}}{a_{ik}} \right) (1 - a_{ik}) \pi_{ik}$$

and the fact that the probabilities sum to one, for $i = 1, \dots, N$.

In one embodiment the covariate matrix X with rows x_i^T can be replaced by a matrix K with ij^{th} element k_{ij} and $k_{ij} = \kappa(x_i - x_j)$ for some kernel function κ . This matrix can also be augmented with a vector of ones. Some example kernels are given in Table 1 below, see Evgeniou et al(1999).

Kernel function	Formula for $\kappa(x - y)$
Gaussian radial basis function	$\exp(- x - y ^2 / a)$, $a > 0$
Inverse multiquadric	$(x - y ^2 + c^2)^{-1/2}$
multiquadric	$(x - y ^2 + c^2)^{1/2}$
Thin plate splines	$ x - y ^{2n+1}$ $ x - y ^{2n} \ln(x - y)$
Multi layer perceptron	$\tanh(x \cdot y - \theta)$, for suitable θ
Ploynomial of degree d	$(1 + x \cdot y)^d$
B splines	$B_{2n+1}(x - y)$
Trigonometric polynomials	$\sin((d + 1/2)(x - y)) / \sin((x - y)/2)$

Table 1: Examples of kernel functions

In Table 1 the last two kernels are preferably one dimensional i.e. for the case when X has only one column.

Multivariate versions can be derived from products of these kernel functions. The definition of B_{2n+1} can be found in De Boor(1978). Use of a kernel function results in mean values

- 47 -

which are smooth (as opposed to transforms of linear) functions of the covariates X . Such models may give a substantially better fit to the data.

- 5 A third embodiment relating to generalised linear models will now be described.

C. Generalised Linear Models

- 10 The method of this embodiment utilises the training samples in order to identify a subset of components which can predict the characteristic of a sample. Subsequently, knowledge of the subset of components can be used for tests, for example clinical tests to predict unknown values of the
15 characteristic of interest. For example, a subset of components of a DNA microarray may be used to predict a clinically relevant characteristic such as, for example, a blood glucose level, a white blood cell count, the size of a tumour, tumour growth rate or survival time.

- 20 In this way, the present invention identifies preferably a relatively small number of components which can be used to predict a characteristic for a particular sample. The selected components are "predictive" for that
25 characteristic. By appropriate choice of hyperparameters in the hyper prior the algorithm can be made to select subsets of varying sizes. Essentially, from all the data which is generated from the system, the method of the present invention enables identification of a small number of
30 components which can be used to predict a particular characteristic. Once those components have been identified, by this method, the components can be used in future to predict the characteristic for new samples. The method of the present invention preferably utilises a statistical
35 method to eliminate components that are not required to correctly predict the characteristic for the sample.

- 48 -

The inventors have found that component weights of a linear combination of components of data generated from the training samples can be estimated in such a way as to eliminate the components that are not required to predict a characteristic for a training sample. The result is that a subset of components are identified which can correctly predict the characteristic for samples in the training set. The method of the present invention thus permits identification from a large amount of data a relatively small number of components which are capable of correctly predicting a characteristic for a training sample, for example, a quantity of interest.

The characteristic may be any characteristic of interest. In one embodiment, the characteristic is a quantity or measure. In another embodiment, they may be the index number of a group, where the samples are grouped into two sample groups (or "classes") based on a pre-determined classification. The classification may be any desired classification by which the training samples are to be grouped. For example, the classification may be whether the training samples are from a leukemia cell or a healthy cell, or that the training samples are obtained from the blood of patients having or not having a certain condition, or that the training samples are from a cell from one of several types of cancer as compared to a normal cell. In another embodiment the characteristic may be a censored survival time, indicating that particular patients have survived for at least a given number of days. In other embodiments the quantity may be any continuously variable characteristic of the sample which is capable of measurement, for example blood pressure.

In one embodiment, the data may be a quantity y_i , where $i \in \{1, \dots, N\}$. We write the $n \times 1$ vector with elements y_i as y . We define a $p \times 1$ parameter vector β of component weights (many of which are expected to be zero), and a $q \times 1$ vector of parameters ϕ (not expected to be zero). Note that q could be

WO 2004/104856

PCT/AU2004/000696

- 49 -

zero (i.e. the set of parameters not expected to be zero may be empty).

In one embodiment, the input data is organised into an
 5 $n \times p$ data matrix $X = (x_{ij})$ with n test training samples and p components. Typically, p will be much greater than n .

In another embodiment, data matrix X may be replaced by an n
 $n \times n$ kernel matrix K to obtain smooth functions of X as
 10 predictors instead of linear predictors. An example of the kernel matrix K is $k_{ij} = \exp(-0.5 * (x_i - x_j)^t (x_i - x_j) / \sigma^2)$ where the subscript on x refers to a row number in the matrix X . Ideally, subsets of the columns of K are selected which give sparse representations of these smooth functions.

15 Typically, as discussed above, the component weights are estimated in a manner which takes into account the apriori assumption that most of the component weights are zero.

20 In one embodiment, the prior specified for the component weights is of the form:

$$p(\beta) = \int_{v^2} p(\beta | v^2) p(v^2) dv^2 \quad (C1)$$

wherein v is a $p \times 1$ vector of hyperparameters, and where
 25 $p(\beta | v^2)$ is $N(0, \text{diag}\{v^2\})$ and $p(v^2)$ is some hyperprior distribution for v^2 .

A suitable form of hyperprior is, $p(v^2) \propto \prod_{i=1}^p p(v_i^2)$. Jeffreys

In another embodiment, the hyperprior $p(v^2)$ is such that
 30 each $t_i^2 = 1/v_i^2$ has an independent gamma distribution.

In another embodiment, the hyperprior $p(v^2)$ is such that each v_i^2 has an independent gamma distribution.

- 50 -

Preferably, an uninformative prior for ϕ is specified.

The likelihood function is defined from a model for the
5 distribution of the data. Preferably, in general, the
likelihood function is any suitable likelihood function. For
example, the likelihood function $L(y|\beta, \phi)$ may be, but not
restricted to, of the form appropriate for a generalised
linear model (GLM), such as for example, that described by
10 Nelder and Wedderburn (1972). Preferably, in this case, the
likelihood function is of the form:

$$L = \log p(y|\beta, \phi) = \sum_{i=1}^N \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} \quad (C2)$$

where $y = (y_1, \dots, y_n)^T$ and $a_i(\phi) = \phi / w_i$ with the w_i being a
15 fixed set of known weights and ϕ a single scale parameter.

Preferably, the likelihood function is specified as follows:
We have

$$\begin{aligned} E\{y_i\} &= b'(\theta_i) \\ 20 \quad \text{Var}\{y\} &= b''(\theta_i) a_i(\phi) = \tau_i^2 a_i(\phi) \end{aligned} \quad (C3)$$

Each observation has a set of covariates x_i and a linear
predictor $\eta_i = x_i^T \beta$. The relationship between the mean of
the i^{th} observation and its linear predictor is given by the
link function $\eta_i = g(\mu_i) = g(b'(\theta_i))$. The inverse of the
25 link is denoted by h , i.e
 $\mu_i = b'(\theta_i) = h(\eta_i)$.

In addition to the scale parameter, a generalised linear
model may be specified by four components:
30

- the likelihood or (scaled) deviance function,
- the link function
- the derivative of the link function
- the variance function.

Some common examples of generalised linear models are given in the table below.

Distribution	Link function $g(\mu)$	Derivative of link function	Variance function	Scale parameter
Gaussian	μ	1	1	Yes
Binomial	$\log(\mu/(1-\mu))$	$1/(\mu(1-\mu))$	$\mu(1-\mu)/n$	No
Poisson	$\log(\mu)$	$1/\mu$	μ	No
Gamma	$1/\mu$	$-1/\mu^2$	μ^2	Yes
Inverse Gaussian	$1/\mu^2$	$-2/\mu^3$	μ^3	Yes

5 In another embodiment, a quasi likelihood model is specified wherein only the link function and variance function are defined. In some instances, such specification results in the models in the table above. In other instances, no distribution is specified.

10

In one embodiment, the posterior distribution of β ϕ and v given y is estimated using:

$$p(\beta \phi v | y) \propto L(y | \beta \phi) p(\beta | v) p(v) \quad (C4)$$

15

wherein $L(y | \beta \phi)$ is the likelihood function.

20

In one embodiment, v may be treated as a vector of missing data and an iterative procedure used to maximise equation (C4) to produce maximum a posteriori estimates of β . The prior of equation (C1) is such that the maximum a posteriori estimates will tend to be sparse i.e. if a large number of parameters are redundant, many components of β will be zero.

25

As stated above, the component weights which maximise the posterior distribution may be determined using an iterative procedure. Preferable, the iterative procedure for

maximising the posterior distribution of the components and component weights is an EM algorithm comprising an E step and an M step, such as, for example, that described in Dempster et al, 1977.

5

In conducting the EM algorithm, the E step preferably comprises the step of computing terms of the form

$$\begin{aligned} P &= \sum_{i=1}^p E\{\beta_i^2 / \nu_i^2 \mid \hat{\beta}_i\} \\ &= \sum_{i=1}^p \beta_i^2 / \hat{d}_i^2 \end{aligned} \quad (C4a)$$

where $\hat{d}_i = d_i(\hat{\beta}_i) = E\{1/\nu_i^2 \mid \hat{\beta}_i\}^{-0.5}$ and for convenience we define
 10 $\hat{d}_i = 1/\hat{d}_i = 0$ if $\hat{\beta}_i = 0$. In the following we write $\hat{d} = d(\hat{\beta}) = (\hat{d}_1, \hat{d}_2, \dots, \hat{d}_p)^T$. In a similar manner we define for example $d(\beta^{(n)})$ and $d(\gamma^{(n)}) = P_n^T d(P_n \gamma^{(n)})$ where $\beta^{(n)} = P_n \gamma^{(n)}$ and P_n is obtained from the p by p identity matrix by omitting columns j for which $\beta_j^{(n)} = 0$.

15 Preferably, equation (C4a) is computed by calculating the conditional expected value of $t_i^2 = 1/\nu_i^2$ when $p(\beta_i \mid \nu_i^2)$ is $N(0, \nu_i^2)$ and $p(\nu_i^2)$ has a specified prior distribution. Specific examples and formulae will be given later.

20 In a general embodiment, appropriate for any suitable likelihood function, the EM algorithm comprises the steps:

(a) Selecting a hyperprior and values for its parameters. Initialising the algorithm by setting $n=0$, $S_0 = \{1, 2, \dots, p\}$, initialise $\varphi^{(0)}$, β^* and applying a value for ϵ , such as for
 25 example $\epsilon = 10^{-5}$;

(b) Defining

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases} \quad (C5)$$

and let P_n be a matrix of zeroes and ones such
 30 that the nonzero elements $\gamma^{(n)}$ of $\beta^{(n)}$ satisfy

- 53 -

$$\begin{aligned}\gamma^{(n)} &= P_n^T \beta^{(n)} \quad , \quad \beta^{(n)} = P_n \gamma^{(n)} \\ \gamma &= P_n^T \beta \quad , \quad \beta = P_n \gamma\end{aligned}$$

(c) performing an estimation (E) step by calculating the conditional expected value of the posterior distribution of component weights using the function:

$$\begin{aligned}Q(\beta | \beta^{(n)}, \phi^{(n)}) &= E\{\log p(\beta, \phi, v | y) | y, \beta^{(n)}, \phi^{(n)}\} \\ &= L(y | \beta, \phi^{(n)}) - 0.5 \beta^T \Delta(d(\beta^{(n)}))^{-2} \beta\end{aligned}\quad (C6)$$

where L is the log likelihood function of y. Using $\beta = P_n \gamma$ and $d(\gamma^{(n)})$ as defined in (C4a), (C6) can be written as

$$Q(\gamma | \gamma^{(n)}, \phi^{(n)}) = L(y | P_n \gamma, \phi^{(n)}) - 0.5 \gamma^T \Delta(d(\gamma^{(n)}))^{-2} \gamma \quad (C7)$$

(d) performing a maximisation (M) step by applying an iterative procedure to maximise Q as a function of γ whereby $\gamma_0 = \gamma^{(n)}$ and for $r=0,1,2$, $\gamma_{r+1} = \gamma_r + \alpha_r \delta_r$ and where α_r is chosen by a line search algorithm to ensure

$$Q(\gamma_{r+1} | \gamma^{(n)}, \phi^{(n)}) > Q(\gamma_r | \gamma^{(n)}, \phi^{(n)}), \text{ and}$$

$$\delta_r = \Delta(d(\gamma^{(n)})) [-\Delta(d(\gamma^{(n)})) \frac{\partial^2 L}{\partial^2 \gamma_r} \Delta(d(\gamma^{(n)})) + I]^{-1} (\Delta(d(\gamma^{(n)})) (\frac{\partial L}{\partial \gamma_r} - \frac{\gamma_r}{d(\gamma^{(n)})})) \quad (C8)$$

where:

$$d(\gamma^{(n)}) = P_n^T d(P_n \gamma^{(n)}) \text{ as in (C4a); and}$$

$$\frac{\partial L}{\partial \gamma_r} = P_n^T \frac{\partial L}{\partial \beta_r} \quad , \quad \frac{\partial^2 L}{\partial^2 \gamma_r} = P_n^T \frac{\partial^2 L}{\partial^2 \beta_r} P_n$$

$$\text{for } \beta_r = P_n \gamma_r .$$

- 54 -

- (e) Let γ^* be the value of γ_r when some convergence criterion is satisfied, for example, $|| \gamma_r - \gamma_{r+1} || < \varepsilon$ (for example 10^{-5});
- 5 (f) Defining $\beta^* = P_n \gamma^*$, $S_{n+1} = \{i: |\beta_i| > \max_j (|\beta_j| * \varepsilon_1)\}$ where ε_1 is a small constant, for example $1e-5$.
- (g) Set $n=n+1$ and choose $\varphi^{(n+1)} = \varphi^{(n)} + \kappa_n (\varphi^* - \varphi^{(n)})$ where φ^* satisfies $\frac{\partial}{\partial \phi} L(y | P_n \gamma^*, \phi) = 0$ and κ_n is a
- 10 damping factor such that $0 < \kappa_n \leq 1$; and
- (h) Check convergence. If $|| \gamma^* - \gamma^{(n)} || < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step (b) above.

15

In another embodiment, $t_i^2 = 1/v_i^2$ has an independent gamma distribution with scale parameter $b > 0$ and shape parameter $k > 0$ so that the density of t_i^2 is

20
$$\gamma(t_i^2, b, k) = b^{-1} (t_i^2/b)^{k-1} \exp(-t_i^2/b) / \Gamma(k)$$

It can be shown that

25
$$E\{t^2 | \beta\} = (2k+1)/(2/b + \beta^2)$$

as follows:

Define

30
$$I(p, b, k) = \int_0^\infty (t^2)^p t \exp(-0.5\beta^2 t^2) \gamma(t^2, b, k) dt^2$$

then

$$I(p, b, k) = b^{p+0.5} \{\Gamma(p+k+0.5)/\Gamma(k)\} (1+0.5b\beta^2)^{-(p+k+0.5)}$$

WO 2004/104856

PCT/AU2004/000696

- 55 -

ProofLet $s = \beta^2 / 2$ then

$$I(p,b,k) = b^{p+0.5} \int_0^{\infty} (t^2/b)^{p+0.5} \exp(-st^2) \gamma(t^2, b, k) dt^2$$

5 Now using the substitution $u = t^2 / b$ we get

$$I(p,b,k) = b^{p+0.5} \int_0^{\infty} (u)^{p+0.5} \exp(-sub) \gamma(u, 1, k) du$$

Now let $s' = bs$ and substitute the expression for $\gamma(u, 1, k)$. This gives

10

$$I(p,b,k) = b^{p+0.5} \int_0^{\infty} \exp(-(s'+1)u) u^{p+k+0.5-1} du / \Gamma(k)$$

Looking up a table of Laplace transforms, eg Abramowitz and Stegun, then gives the result.

15 The conditional expectation follows from

$$\begin{aligned} E\{t^2 | \beta\} &= I(1, b, k) / I(0, b, k) \\ &= (2k+1) / (2/b + \beta^2) \end{aligned}$$

As k tends to zero and b tends to infinity we get the
20 equivalent result using Jeffreys prior. For example, for
 $k=0.005$ and $b=2 \times 10^5$

$$E\{t^2 | \beta\} = (1.01) / (10^{-5} + \beta^2)$$

Hence we can get arbitrarily close to the algorithm with a
25 Jeffreys hyperprior with this proper prior.

In another embodiment, v_i^2 has an independent gamma distribution with scale parameter $b > 0$ and shape parameter $k > 0$. It can be shown that

$$\begin{aligned}
 E\{\nu_i^{-2}|\beta_i\} &= \frac{\int_0^{\infty} u^{k-3/2-1} \exp(-(\lambda_i/u+u)) du}{b \int_0^{\infty} u^{k-1/2-1} \exp(-(\lambda_i/u+u)) du} \\
 &= \sqrt{\frac{2}{b}} \frac{1}{|\beta_i|} \frac{K_{3/2-k}(2\sqrt{\lambda_i})}{K_{1/2-k}(2\sqrt{\lambda_i})} \\
 &= \frac{1}{|\beta_i|^2} \frac{(2\sqrt{\lambda_i})K_{3/2-k}(2\sqrt{\lambda_i})}{K_{1/2-k}(2\sqrt{\lambda_i})}
 \end{aligned} \tag{C9}$$

where $\lambda_i = \beta_i^2/2b$ and K denotes a modified Bessel function, which can be shown as follows:

For $k=1$ in equation (c9)

5

$$E\{\nu_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|)$$

For $K=0.5$ in equation (C9)

10

$$E\{\nu_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|) \{K_1(2\sqrt{\lambda_i})/K_0(2\sqrt{\lambda_i})\}$$

or equivalently

15

$$E\{\nu_i^{-2}|\beta_i\} = (1/|\beta_i|^2) \{2\sqrt{\lambda_i}K_1(2\sqrt{\lambda_i})/K_0(2\sqrt{\lambda_i})\}$$

Proof

From the definition of the conditional expectation, writing $\lambda_i = \beta_i^2/2b$, we get

20

$$E\{\nu_i^{-2}|\beta_i\} = \frac{\int_0^{\infty} \nu_i^{-2} \nu_i^{-1} \exp(-\lambda_i \nu_i^{-2}) b^{-1} (\nu_i^{-2}/b)^{k-1} \exp(\nu_i^{-2}/b) d\nu_i^{-2}}{\int_0^{\infty} \nu_i^{-1} \exp(-\lambda_i \nu_i^{-2}) b^{-1} (\nu_i^{-2}/b)^{k-1} \exp(\nu_i^{-2}/b) d\nu_i^{-2}}$$

Rearranging, simplifying and making the substitution $u = \nu_i^2/b$ gives A.1

The integrals in A.1 can be evaluated by using the result

25

WO 2004/104856

PCT/AU2004/000696

- 57 -

$$\int_0^{\infty} x^{-b-1} \exp\left[-\left(x + \frac{a^2}{x}\right)\right] dx = \frac{2}{a^b} K_b(2a)$$

where K denotes a modified Bessel function, see Watson(1966).

5 Examples of members of this class are $k=1$ in which case

$$E\{\nu_i^{-2} | \beta_i\} = \sqrt{2/b} (1/|\beta_i|)$$

which corresponds to the prior used in the Lasso technique,
10 Tibshirani(1996). See also Figueiredo(2001).

The case $k=0.5$ gives

$$E\{\nu_i^{-2} | \beta_i\} = \sqrt{2/b} (1/|\beta_i|) \{K_1(2\sqrt{\lambda_i})/K_0(2\sqrt{\lambda_i})\}$$

15

or equivalently

$$E\{\nu_i^{-2} | \beta_i\} = (1/|\beta_i|^2) \{2\sqrt{\lambda_i} K_1(2\sqrt{\lambda_i})/K_0(2\sqrt{\lambda_i})\}$$

20 where K_0 and K_1 are modified Bessel functions, see Abramowitz and Stegun(1970). Polynomial approximations for evaluating these Bessel functions can be found in Abramowitz and Stegun(1970, p379). Details of the above calculations are given in the Appendix.

25

The expressions above demonstrate the connection with the Lasso model and the Jeffreys prior model.

It will be appreciated by those skilled in the art that as k
30 tends to zero and b tends to infinity the prior tends to a Jeffreys improper prior.

In one embodiment, the priors with $0 < k \leq 1$ and $b > 0$ form a class of priors which might be interpreted as penalising non
35 zero coefficients in a manner which is between the Lasso prior and the original specification using Jeffreys prior.

For generalised linear models the expected value $E\{\frac{\partial^2 L}{\partial^2 \gamma_r}\}$ may be calculated as follows. Beginning with

where X is the N by p matrix with i^{th} row x_i^T and

we get

20 Equations (C10) and (C11) can be written as

where $V = \Delta(a_i(\varphi) \tau_i^2 (\frac{\partial \eta_i}{\partial \mu_i})^2)$.

WO 2004/104856

PCT/AU2004/000696

- 59 -

Preferably, for generalised linear models, the EM algorithm comprises the steps:

(a) Choose a hyper prior and its parameters.

5 Initialising the algorithm by setting $n=0$, $S_0 = \{1, 2, \dots, p\}$, $\phi^{(0)}$, applying a value for ϵ , such as for example $\epsilon = 10^{-5}$, and

If $p \leq N$ compute initial values β^* by

$$\beta^* = (X^T X + \lambda I)^{-1} X^T g(y + \zeta) \quad (C14)$$

and if $p > N$ compute initial values β^* by

10
$$\beta^* = \frac{1}{\lambda} (I - X^T (X X^T + \lambda I)^{-1} X) X^T g(y + \zeta) \quad (C15)$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$ and ζ is small and chosen so that the link function is well defined at $y + \zeta$.

(b) Defining

15

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

and let P_n be a matrix of zeroes and ones such that the nonzero elements $\gamma^{(n)}$ of $\beta^{(n)}$ satisfy

20

$$\begin{aligned} \gamma^{(n)} &= P_n^T \beta^{(n)}, & \beta^{(n)} &= P_n \gamma^{(n)} \\ \gamma &= P_n^T \beta, & \beta &= P_n \gamma \end{aligned}$$

(c) performing an estimation (E) step by calculating
25 the conditional expected value of the posterior distribution of component weights using the function:

$$\begin{aligned} Q(\beta | \beta^{(n)}, \phi^{(n)}) &= E\{\log p(\beta, \phi, v | y) | y, \beta^{(n)}, \phi^{(n)}\} \\ &= L(y | \beta, \phi^{(n)}) - 0.5 \beta^T \Delta(d(\beta^{(n)}))^{-2} \beta \end{aligned} \quad (C16)$$

30

WO 2004/104856

PCT/AU2004/000696

- 60 -

where L is the log likelihood function of y .

Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ (C16) can be written as

$$Q(\gamma | \gamma^{(n)}, \phi^{(n)}) = L(y | P_n \gamma, \phi^{(n)}) - 0.5 \gamma^T \Delta(d(\gamma^{(n)}))^{-2} \gamma \quad (C17)$$

- (d) performing a maximisation (M) step by applying an iterative procedure, for example a Newton Raphson iteration, to maximise Q as a function of γ whereby $\gamma_0 = \gamma^{(n)}$ and for $r=0, 1, 2, \dots$ $\gamma_{r+1} = \gamma_r + \alpha_r \delta_r$ where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1} | \gamma^{(n)}, \phi^{(n)}) > Q(\gamma_r | \gamma^{(n)}, \phi^{(n)})$, and for $p \leq N$ use

$$\delta_r = \Delta(d(\gamma^{(n)})) [Y_n^T V_r^{-1} Y_n + I]^{-1} (Y_n^T V_r^{-1} z_r - \frac{\gamma_r}{d(\gamma^{(n)})}) \quad (C18)$$

where

$$Y_n = \Delta(d(\gamma^{(n)})) P_n^T X$$

$$V = \Delta(a_i(\phi) \tau_i^2 (\frac{\partial \eta_i}{\partial \mu_i})^2)$$

$$z = \frac{\partial \eta}{\partial \mu} (y - \mu)$$

and the subscript r denotes that these quantities are evaluated at $\mu = h(X P_n \gamma_r)$.

For $p > N$ use

$$\delta_r = \Delta(d(\gamma^{(n)})) [I - Y_n^T (Y_n Y_n^T + V_r)^{-1} Y_n] (Y_n^T V_r^{-1} z_r - \frac{\gamma_r}{d(\gamma^{(n)})}) \quad (C19)$$

with V_r and z_r defined as before.

- (e) Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g

$$|| \gamma_r - \gamma_{r+1} || < \epsilon \quad (\text{for example } 10^{-5}).$$

- 61 -

(f) Define $\beta^* = P_n \gamma^*$, $S_{n+1} = \{i: |\beta_i| > \max_j (|\beta_j| * \epsilon_1)\}$ where

ϵ_1 is a small constant, say $1e-5$. Set $n=n+1$ and choose $\phi^{n+1} = \phi^n + \kappa_n (\phi^* - \phi^n)$ where ϕ^* satisfies $\frac{\partial}{\partial \phi} L(y | P_n \gamma^*, \phi) = 0$ and κ_n is a damping factor such that

5 $0 < \kappa_n \leq 1$. Note that in some cases the scale parameter is known or this equation can be solve explicitly to get an updating equation for ϕ .

The above embodiments may be extended to incorporate quasi
10 likelihood methods Wedderburn (1974) and McCullagh and Nelder (1983)). In such an embodiment, the same iterative procedure as detailed above will be appropriate, but with L replaced by a quasi likelihood as shown above and, for example, Table 8.1 in McCullagh and Nelder (1983). In one
15 embodiment there is a modified updating method for the scale parameter ϕ . To define these models requires specification of the variance function τ^2 , the link function g and the derivative of the link function $\frac{\partial \eta}{\partial \mu}$. Once these are defined the above algorithm can be applied.

20 In one embodiment for quasi likelihood models, step 5 of the above algorithm is modified so that the scale parameter is updated by calculating

25
$$\phi^{(n+1)} = \frac{1}{N} \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\tau_i^2}$$

where μ and τ are evaluated at $\beta^* = P_n \gamma^*$. Preferably, this updating is performed when the number of parameters s in the model is less than N . A divisor of $N-s$ can be used when s
30 is much less than N .

WO 2004/104856

PCT/AU2004/000696

- 62 -

In another embodiment, for both generalised linear models and Quasi likelihood models the covariate matrix X with rows x_i^T can be replaced by a matrix K with ij^{th} element k_{ij} and $k_{ij} = \kappa(x_i - x_j)$ for some kernel function κ . This matrix can also be augmented with a vector of ones. Some example kernels are given in Table 2 below, see Evgeniou et al(1999).

Kernel function	Formula for $\kappa(x - y)$
Gaussian radial basis function	$\exp(- x - y ^2 / a), a > 0$
Inverse multiquadric	$(x - y ^2 + c^2)^{-1/2}$
Multiquadric	$(x - y ^2 + c^2)^{1/2}$
Thin plate splines	$ x - y ^{2n+1}$ $ x - y ^{2n} \ln(x - y)$
Multi layer perceptron	$\tanh(x \cdot y - \theta)$, for suitable θ
Polynomial of degree d	$(1 + x \cdot y)^d$
B splines	$B_{2n+1}(x - y)$
Trigonometric polynomials	$\sin((d + 1/2)(x - y)) / \sin((x - y)/2)$

10

Table 2: Examples of kernel functions

In Table 2 the last two kernels are one dimensional i.e. for the case when X has only one column. Multivariate versions can be derived from products of these kernel functions. The definition of B_{2n+1} can be found in De Boor(1978). Use of a kernel function in either a generalised linear model or a quasi likelihood model results in mean values which are smooth (as opposed to transforms of linear) functions of the covariates X . Such models may give a substantially better fit to the data.

20

- 63 -

A fourth embodiment relating to a proportional hazards model will now be described.

5 D. Proportional Hazard Models

The method of this embodiment may utilise training samples in order to identify a subset of components which are capable of affecting the probability that a defined event
10 (eg death, recovery) will occur within a certain time period. Training samples are obtained from a system and the time measured from when the training sample is obtained to when the event has occurred. Using a statistical method to associate the time to the event with the data obtained from
15 a plurality of training samples, a subset of components may be identified that are capable of predicting the distribution of the time to the event. Subsequently, knowledge of the subset of components can be used for tests, for example clinical tests to predict for example,
20 statistical features of the time to death or time to relapse of a disease. For example, the data from a subset of components of a system may be obtained from a DNA microarray. This data may be used to predict a clinically relevant event such as, for example, expected or median
25 patient survival times, or to predict onset of certain symptoms, or relapse of a disease.

In this way, the present invention identifies preferably a relatively small number of components which can be used to
30 predict the distribution of the time to an event of a system. The selected components are "predictive" for that time to an event. Essentially, from all the data which is generated from the system, the method of the present invention enables identification of a small number of
35 components which can be used to predict time to an event. Once those components have been identified by this method, the components can be used in future to predict statistical

PCT/AU2004/000696

- 64 -

5 of a system. By appropriate selection of the hyperparameters
in the model some control over the size of the selected
subset can be achieved.

10 the time from obtaining the sample to which the method of
the invention is applied to the time of an event. An event
may be any observable event. When the system is a
biological system, the event may be, for example, time till
failure of a system, time till death, onset of a particular
15 symptom or symptoms, onset or relapse of a condition or
disease, change in phenotype or genotype, change in
biochemistry, change in morphology of an organism or tissue,
change in behaviour.

20 The samples are associated with a particular time to an event from previous times to an event. The times to an event may be times determined from data obtained from, for example, patients in which the time from sampling to death is known, or in other words, "genuine" survival times, and
25 patients in which the only information is that the patients were alive when samples were last obtained, or in other words, "censored" survival times indicating that the particular patient has survived for at least a given number of days.

30 In one embodiment, the input data is organised into an $n \times p$ data matrix $X = (x_{ij})$ with n test training samples and p components. Typically, p will be much greater than n .

35 For example, consider an $N \times p$ data matrix $X = (x_{ij})$ from, for example, a microarray experiment, with N individuals (or samples) and the same p genes for each individual.

WO 2004/104856

PCT/AU2004/000696

- 65 -

Preferably, there is associated with each individual i ($i=1,2,\dots,N$) a variable y_i ($y_i \geq 0$) denoting the time to an event, for example, survival time. For each individual there may also be defined a variable that indicates whether that individual's survival time is a genuine survival time or a censored survival time. Denote the censor indicators as c_i where

$$c_i = \begin{cases} 1, & \text{if } y_i \text{ is uncensored} \\ 0, & \text{if } y_i \text{ is censored} \end{cases}$$

10

The $N \times 1$ vector with survival times y_i may be written as \underline{y} and the $N \times 1$ vector with censor indicators c_i as \underline{c} .

Typically, as discussed above, the component weights are estimated in a manner which takes into account the a priori assumption that most of the component weights are zero.

15

Preferably, the prior specified for the component weights is of the form

20

$$P(\beta_1, \beta_2, \dots, \beta_n) = \int \prod_{i=1}^N P(\beta_i | \tau_i) P(\tau_i) d\tau \quad (D1)$$

where $\beta_1, \beta_2, \dots, \beta_n$ are component weights, $P(\beta_i | \tau_i)$ is $N(0, \tau_i^2)$ and $P(\tau_i)$ is some hyperprior distribution

25

$$P(\tau) = \prod_{i=1}^n P(\tau_i)$$

that is not a Jeffrey's hyperprior.

In one embodiment, the prior distribution is an inverse gamma prior for τ in which $t_i^2 = 1/\tau_i^2$ has an independent gamma distribution with scale parameter $b > 0$ and shape parameter $k > 0$ so that the density of t_i^2 is

30

WO 2004/104856

PCT/AU2004/000696

- 66 -

$$\gamma(t_i^2, b, k) = b^{-1} (t_i^2/b)^{k-1} \exp(-t_i^2/b) / \Gamma(k) .$$

It can be shown that:

$$5 \quad E\{t^2 | \beta\} = (2k+1)/(2/b + \beta^2) \quad (A)$$

Equation A can be shown as follows:

Define

$$10 \quad I(p, b, k) = \int_0^\infty (t^2)^p t \exp(-0.5\beta^2 t^2) \gamma(t^2, b, k) dt^2$$

then

$$I(p, b, k) = b^{p+0.5} \{\Gamma(p+k+0.5)/\Gamma(k)\} (1+0.5b\beta^2)^{-(p+k+0.5)}$$

15 Proof

Let $s = \beta^2 / 2$ then

$$I(p, b, k) = b^{p+0.5} \int_0^\infty (t^2/b)^{p+0.5} \exp(-st^2) \gamma(t^2, b, k) dt^2$$

Now using the substitution $u = t^2 / b$ we get

20

$$I(p, b, k) = b^{p+0.5} \int_0^\infty (u)^{p+0.5} \exp(-sub) \gamma(u, 1, k) du$$

Now let $s' = bs$ and substitute the expression for $\gamma(u, 1, k)$. This gives

$$I(p, b, k) = b^{p+0.5} \int_0^\infty \exp(-(s'+1)u) u^{p+k+0.5-1} du / \Gamma(k)$$

25 Looking up a table of Laplace transforms, eg Abramowitz and Stegun, then gives the result.

The conditional expectation follows from

$$30 \quad E\{t^2 | \beta\} = I(1, b, k) / I(0, b, k) \\ = (2k+1)/(2/b + \beta^2)$$

WO 2004/104856

PCT/AU2004/000696

- 67 -

As k tends to zero and b tends to infinity, a result equivalent to using Jeffreys prior is obtained. For example, for $k=0.005$ and $b=2 \times 10^5$

5

$$E\{t^2 | \beta\} = (1.01)/(10^{-5} + \beta^2)$$

Hence we can get arbitrarily close to a Jeffery's prior with this proper prior.

10 The modified algorithm for this model has

$$b^{(n)} = E\{t^2 | \beta^{(n)}\}^{-0.5}$$

where the expectation is calculated as above.

15

In yet another embodiment, the prior distribution is a gamma distribution for τ_{ig}^2 . Preferably, the gamma distribution has scale parameter $b > 0$ and shape parameter $k > 0$.

20 It can be shown, that

$$\begin{aligned} E\{\tau_i^{-2} | \beta_i\} &= \frac{\int_0^\infty u^{k-3/2-1} \exp(-(\gamma_i/u+u)) du}{b \int_0^\infty u^{k-1/2-1} \exp(-(\gamma_i/u+u)) du} \\ &= \frac{\sqrt{2}}{b} \frac{1}{|\beta_i|} \frac{K_{3/2-k}(2\sqrt{\gamma_i})}{K_{1/2-k}(2\sqrt{\gamma_i})} \\ &= \frac{1}{|\beta_i|^2} \frac{(2\sqrt{\gamma_i}) K_{3/2-k}(2\sqrt{\gamma_i})}{K_{1/2-k}(2\sqrt{\gamma_i})} \end{aligned}$$

where $\gamma_i = \beta_i^2 / 2b$ and K denotes a modified Bessel function.

25 Some special members of this class are $k=1$ in which case

$$E\{\tau_i^{-2} | \beta_i\} = \sqrt{2/b} (1/|\beta_i|)$$

WO 2004/104856

PCT/AU2004/000696

- 68 -

which corresponds to the prior used in the Lasso technique, Tibshirani(1996). See also Figueiredo(2001).

The case $k=0.5$ gives

5

$$E\{\tau_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|)\{K_1(2\sqrt{\gamma_i})/K_0(2\sqrt{\gamma_i})\}$$

or equivalently

10

$$E\{\tau_i^{-2}|\beta_i\} = (1/|\beta_i|^2)\{2\sqrt{\gamma_i}K_1(2\sqrt{\gamma_i})/K_0(2\sqrt{\gamma_i})\}$$

where K_0 and K_1 are modified Bessel functions, see Abramowitz and Stegun(1970). Polynomial approximations for evaluating these Bessel functions can be found in Abramowitz and Stegun(1970, p379).

15

The expressions above demonstrate the connection with the Lasso model and the Jeffreys prior model.

20 Details of the above calculations are as follows:

For the gamma prior above and $\gamma_i = \beta_i^2/2b$

$$\begin{aligned} E\{\tau_i^{-2}|\beta_i\} &= \frac{\int_0^\infty u^{k-3/2-1} \exp(-(\gamma_i/u+u)) du}{b \int_0^\infty u^{k-1/2-1} \exp(-(\gamma_i/u+u)) du} \\ &= \sqrt{\frac{2}{b}} \frac{1}{|\beta_i|} \frac{K_{3/2+k}(2\sqrt{\gamma_i})}{K_{1/2+k}(2\sqrt{\gamma_i})} \\ &= \frac{1}{|\beta_i|^2} \frac{(2\sqrt{\gamma_i})K_{3/2+k}(2\sqrt{\gamma_i})}{K_{1/2+k}(2\sqrt{\gamma_i})} \end{aligned} \quad (D2)$$

25

where K denotes a modified Bessel function.
For $k=1$ in (D2)

$$E\{\tau_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|)$$

WO 2004/104856

PCT/AU2004/000696

- 69 -

For $K=0.5$ in (D2)

$$E\{\tau_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|)\{K_1(2\sqrt{\gamma_i})/K_0(2\sqrt{\gamma_i})\}$$

5

or equivalently

$$E\{\tau_i^{-2}|\beta_i\} = (1/|\beta_i|^2)\{2\sqrt{\gamma_i}K_1(2\sqrt{\gamma_i})/K_0(2\sqrt{\gamma_i})\}$$

10 Proof

From the definition of the conditional expectation, writing $\gamma_i = \beta_i^2/2b$, we get

$$E\{\tau_i^{-2}|\beta_i\} = \frac{\int_0^\infty \tau_i^{-2} \tau_i^{-1} \exp(-\gamma_i \tau_i^{-2}) b^{-1} (\tau_i^{-2}/b)^{k-1} \exp(\tau_i^{-2}/b) d\tau_i^{-2}}{\int_0^\infty \tau_i^{-1} \exp(-\gamma_i \tau_i^{-2}) b^{-1} (\tau_i^{-2}/b)^{k-1} \exp(\tau_i^{-2}/b) d\tau_i^{-2}}$$

15 Rearranging, simplifying and making the substitution $u = \tau_i^{-2}/b$ gives A.1

The integrals in A.1 can be evaluated by using the result

$$20 \quad \int_0^\infty x^{-b-1} \exp\left[-\left(x + \frac{a^2}{x}\right)\right] dx = \frac{2}{a^b} K_b(2a)$$

where K denotes a modified Bessel function, see Watson(1966).

25 In a particularly preferred embodiment, $p(\tau_i)\alpha 1/\tau_i^2$ is a Jeffreys prior, Kotz and Johnson(1983).

The likelihood function defines a model which fits the data based on the distribution of the data. Preferably, the likelihood function is of the form:

30

$$\text{Log (Partial) Likelihood} = \sum_{i=1}^N g_i(\beta, \varphi; X, y, \varepsilon)$$

WO 2004/104856

PCT/AU2004/000696

- 70 -

where $\underline{\beta}^T = (\beta_1, \beta_2, \dots, \beta_p)$ and $\underline{\varphi}^T = (\varphi_1, \varphi_2, \dots, \varphi_q)$ are the model parameters. The model defined by the likelihood function may be any model for predicting the time to an event of a system.

5

In one embodiment, the model defined by the likelihood is Cox's proportional hazards model. Cox's proportional hazards model was introduced by Cox (1972) and may preferably be used as a regression model for survival data. In Cox's
10 proportional hazards model, $\underline{\beta}'$ is a vector of (explanatory) parameters associated with the components. Preferably, the method of the present invention provides for the parsimonious selection (and estimation) from the parameters $\underline{\beta}' = (\beta_1, \beta_2, \dots, \beta_p)$ for Cox's proportional hazards model given
15 the data X , y and c .

Application of Cox's proportional hazards model can be problematic in the circumstance where different data is obtained from a system for the same survival times, or in
20 other words, tied survival times. Tied survival times may be subjected to a pre-processing step that leads to unique survival times. The pre-processing proposed simplifies the ensuing code as it avoids concerns about tied survival times in the subsequent application of Cox's proportional hazards
25 model.

The pre-processing of the survival times applies by adding an extremely small amount of insignificant random noise. Preferably, the procedure is to take sets of tied times and
30 add to each tied time within a set of tied times a random amount that is drawn from a normal distribution that has zero mean and variance proportional to the smallest non-zero distance between sorted survival times. Such pre-processing achieves an elimination of tied times without imposing a
35 draconian perturbation of the survival times.

- 71 -

The pre-processing generates distinct survival times.

Preferably, these times may be ordered in increasing

magnitude denoted as $\underline{t} = (t_{(1)}, t_{(2)}, \dots, t_{(N)})$, $t_{(i+1)} > t_{(i)}$.

Denote by Z the $N \times p$ matrix that is the re-arrangement
 5 of the rows of X where the ordering of the rows of
 Z corresponds to the ordering induced by the ordering of \underline{t} ;
 also denote by Z_j the j^{th} row of the matrix Z . Let d be the
 result of ordering c with the same permutation required to
 order \underline{t} .

10

After pre-processing for tied survival times is taken
 into account and reference is made to standard texts on
 survival data analysis (eg Cox and Oakes, 1984), the
 likelihood function for the proportional hazards model may
 15 preferably be written as

$$l(\underline{t} | \underline{\beta}) = \prod_{j=1}^N \left(\frac{\exp(Z_j \underline{\beta})}{\sum_{i \in \mathfrak{R}_j} \exp(Z_i \underline{\beta})} \right)^{d_j} \quad (\text{D3})$$

where $\underline{\beta}^T = (\beta_1, \beta_2, \dots, \beta_n)$, Z_j = the j^{th} row of Z , and

$\mathfrak{R}_j = \{i : i = j, j+1, \dots, N\}$ = the risk set at the j^{th} ordered event

20 time $t_{(j)}$.

The logarithm of the likelihood (ie $L = \log(l)$) may
 preferably be written as

$$\begin{aligned} L(\underline{t} | \underline{\beta}) &= \sum_{i=1}^N d_i \left(Z_i \underline{\beta} - \log \left(\sum_{j \in \mathfrak{R}_i} \exp(Z_j \underline{\beta}) \right) \right) \\ &= \sum_{i=1}^N d_i \left(Z_i \underline{\beta} - \log \left(\sum_{j=1}^N \zeta_{i,j} \exp(Z_j \underline{\beta}) \right) \right), \end{aligned} \quad (\text{D4})$$

where

$$\zeta_{i,j} = \begin{cases} 0, & \text{if } j < i \\ 1, & \text{if } j \geq i \end{cases}$$

WO 2004/104856

PCT/AU2004/000696

- 72 -

Notice that the model is non-parametric in that the parametric form of the survival distribution is not specified - preferably only the ordinal property of the survival times are used (in the determination of the risk sets). As this is a non-parametric case φ is not required (ie $q=0$).

In another embodiment of the method of the invention, the model defined by the likelihood function is a Parametric survival model. Preferably, in a parametric survival model, β' is a vector of (explanatory) parameters associated with the components, and φ' is a vector of parameters associated with the functional form of the survival density function.

Preferably, the method of the invention provides for the parsimonious selection (and estimation) from the parameters β' and the estimation of $\varphi' = (\varphi_1, \varphi_2, \dots, \varphi_q)$ for parametric survival models given the data X , y and c .

In applying a parametric survival model, the survival times do not require pre-processing and are denoted as y . The parametric survival model is applied as follows: Denote by $f(y; \varphi, \beta, X)$ the parametric density function of the survival time, denote its survival function by

$$S(y; \varphi, \beta, X) = \int_y^{\infty} f(u; \varphi, \beta, X) du \text{ where } \varphi \text{ are the parameters relevant}$$

to the parametric form of the density function and β, X are as defined above. The hazard function is defined as

$$h(y_i; \varphi, \beta, X) = f(y_i; \varphi, \beta, X) / S(y_i; \varphi, \beta, X).$$

Preferably, the generic formulation of the log-likelihood function, taking censored data into account, is

$$L = \sum_{i=1}^N \left\{ c_i \log \left(f(y_i; \underline{\varphi}, \underline{\beta}, X) \right) + (1 - c_i) \log \left(S(y_i; \underline{\varphi}, \underline{\beta}, X) \right) \right\}$$

Reference to standard texts on analysis of survival time data via parametric regression survival models reveals a collection of survival time distributions that may be used.

- 5 Survival distributions that may be used include, for example, the Weibull, Exponential or Extreme Value distributions.

If the hazard function may be written as

10 $h(y_i; \underline{\varphi}, \underline{\beta}, X) = \lambda(y_i; \underline{\varphi}) \exp(X_i \underline{\beta})$ then $S(y_i; \underline{\varphi}, \underline{\beta}, X) = \exp \left(-\Lambda(y_i; \underline{\varphi}) e^{X_i \underline{\beta}} \right)$ and

$$f(y_i; \underline{\varphi}, \underline{\beta}, X) = \lambda(y_i; \underline{\varphi}) \exp \left(X_i \underline{\beta} - \Lambda(y_i) e^{X_i \underline{\beta}} \right) \text{ where } \Lambda(y_i; \underline{\varphi}) = \int_{-\infty}^{y_i} \lambda(u; \underline{\varphi}) du$$

is the integrated hazard function and $\lambda(y_i; \underline{\varphi}) = \frac{d\Lambda(y_i; \underline{\varphi})}{dy_i}$; X_i is the i^{th} row of X .

- 15 The Weibull, Exponential and Extreme Value distributions have density and hazard functions that may be written in the form of those presented in the paragraph immediately above.
- 20 The application detailed relies in part on an algorithm of Aitken and Clayton (1980) however it permits the user to specify any parametric underlying hazard function.

- Following from Aitkin and Clayton (1980) a preferred
- 25 likelihood function which models a parametric survival model is:

$$L = \sum_{i=1}^N \left\{ c_i \log(\mu_i) - \mu_i + c_i \left(\log \left(\frac{\lambda(y_i)}{\Lambda(y_i; \underline{\varphi})} \right) \right) \right\} \quad (D5)$$

where $\mu_i = \Lambda(y_i; \underline{\varphi}) \exp(X_i \underline{\beta})$. Aitkin and Clayton (1980) note that a consequence of equation (11) is that the c_i 's may be

WO 2004/104856

PCT/AU2004/000696

- 74 -

treated as Poisson variates with means μ_i and that the last term in equation (11) does not depend on $\underline{\beta}$ (although it depends on $\underline{\varphi}$).

- 5 Preferably, the posterior distribution of $\underline{\beta}$, $\underline{\varphi}$ and $\underline{\tau}$ given \underline{y} is

$$P(\underline{\beta}, \underline{\varphi}, \underline{\tau} | \underline{y}) \propto l(\underline{y} | \underline{\beta}, \underline{\varphi}, \underline{\tau}) P(\underline{\beta} | \underline{\tau}) P(\underline{\tau}) \quad (D6)$$

wherein $l(\underline{y} | \underline{\beta}, \underline{\varphi}, \underline{\tau})$ is the likelihood function.

10

In one embodiment, $\underline{\tau}$ may be treated as a vector of missing data and an iterative procedure used to maximise equation (D6) to produce a posteriori estimates of $\underline{\beta}$. The prior of equation (D1) is such that the maximum a posteriori estimates will tend to be sparse i.e. if a large number of parameters are redundant, many components of $\underline{\beta}$ will be zero.

15

- 20 Because a prior expectation exists that many components of $\underline{\beta}^i$ are zero, the estimation may be performed in such a way that most of the estimated β_i 's are zero and the remaining non-zero estimates provide an adequate explanation of the survival times.

- 25 In the context of microarray data this exercise translates to identifying a parsimonious set of genes that provide an adequate explanation for the event times.

- 30 As stated above, the component weights which maximise the posterior distribution may be determined using an iterative procedure. Preferable, the iterative procedure for maximising the posterior distribution of the components and component weights is an EM algorithm, such as, for example, that described in Dempster et al, 1977.

35

- 75 -

If the E step of the EM algorithm is examined, from (D6) ignoring terms not involving beta, it is necessary to compute

$$\begin{aligned} & \sum_{i=1}^n E\{\beta_i^2/\tau_i^2 \mid \hat{\beta}_i\} \\ &= \sum_{i=1}^n \beta_i^2/\hat{d}_i^2 \end{aligned} \quad (D7)$$

5 where $\hat{d}_i = d_i(\hat{\beta}_i) = E\{1/v_i^2 \mid \hat{\beta}_i\}^{-0.5}$ and for convenience we define $\hat{d}_i = 1/\hat{d}_i = 0$ if $\hat{\beta}_i = 0$. In the following we write $\hat{d} = d(\hat{\beta}) = (\hat{d}_1, \hat{d}_2, \dots, \hat{d}_n)^T$. In a similar manner we define for example $d(\beta^{(n)})$ and $d(\gamma^{(n)}) = P_n^T d(P_n \gamma^{(n)})$ where $\beta^{(n)} = P_n \gamma^{(n)}$ and P_n is obtained from the p by p identity matrix by omitting columns j for which $\beta_j^{(n)} = 0$.

10

Hence to do the E step requires the calculation of the conditional expected value of $t_i^2 = 1/\tau_i^2$ when $p(\beta_i \mid \tau_i^2)$ is $N(0, \tau_i^2)$ and $p(\tau_i^2)$ has a specified prior distribution as discussed above.

15

In one embodiment, the EM algorithm comprises the steps:

1. Choose the hyperprior and values for its parameters, namely b and k. Initialising the algorithm by setting $n=0$,

20 $S_0 = \{1, 2, \dots, p\}$, initialise $\underline{\beta}^{(0)} = \underline{\beta}^*$, $\underline{\varphi}^{(0)}$,

2. Defining

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

25

and let P_n be a matrix of zeroes and ones such that the nonzero elements $\underline{\gamma}^{(n)}$ of $\underline{\beta}^{(n)}$ satisfy

$$\begin{aligned} \underline{\gamma}^{(n)} &= P_n^T \underline{\beta}^{(n)}, & \underline{\beta}^{(n)} &= P_n \underline{\gamma}^{(n)} \\ \underline{\gamma} &= P_n^T \underline{\beta}, & \underline{\beta} &= P_n \underline{\gamma} \end{aligned} \quad (D8)$$

This may be performed using the function:

$$\begin{aligned} Q(\underline{\beta} | \underline{\beta}^{(n)}, \varphi^{(n)}) &= E\left\{\log\left(P(\underline{\beta}, \underline{\varphi}, \tau | \underline{y})\right) | \underline{y}, \underline{\beta}^{(n)}, \varphi^{(n)}\right\} \\ &= L(\underline{y} | \underline{\beta}, \varphi^{(n)}) - \frac{1}{2} \sum_{i=1}^p \left(\frac{\beta_i}{d_i(\beta^{(n)})} \right)^2 \end{aligned} \quad (D9)$$

where L is the log likelihood function of \underline{y} . Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ we have

$$10 \quad Q(\gamma|\gamma^{(n)}, \varphi^{(n)}) = L(t|P_n\gamma, \varphi^{(n)}) - \frac{1}{2}\gamma^T \Delta(d(\gamma^{(n)}))\gamma \quad (\text{D10})$$

4. Performing the maximisation step. This may be performed using Newton Raphson iterations as follows:

Set $\gamma_0 = \gamma^{(r)}$ and for $r=0, 1, 2, \dots$ $\gamma_{r+1} = \gamma_r + \alpha_r \delta_r$ where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1} | \gamma^{(n)}, \varphi^{(n)}) >$

$$15 \quad Q(\gamma_r \mid \gamma^{(n)}, \varphi^{(n)}), \quad \text{and}$$

$$\underline{\delta}_r = \Delta(d(\underline{y}^{(n)}))[-\Delta(d(\underline{y}^{(n)}))\frac{\partial^2 L}{\partial^2 \gamma_r}\Delta(d(\underline{y}^{(n)})) + I]^{-1}(\frac{\partial L}{\partial \gamma_r} - \frac{\gamma_r}{d(\underline{y}^{(n)})})$$

$$\text{where } \frac{\partial L}{\partial \gamma_r} = P_n^T \frac{\partial L}{\partial \beta_r}, \quad \frac{\partial^2 L}{\partial^2 \gamma_r} = P_n^T \frac{\partial^2 L}{\partial^2 \beta_r} P_n \quad \text{for } \beta_r = P_n \gamma_r \quad (\text{D11})$$

Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g. $\|\gamma_r - \gamma_{r+1}\| < \varepsilon$ (for example $\varepsilon = 10^{-5}$)

20 5. Define $\beta_n^* = P_n \gamma_n^*$, $S_n = \left\{ i : |\beta_i| > \varepsilon_1 \max_j |\beta_j| \right\}$ where ε_1 is a small

constant, say 10^{-5} . Set $n=n+1$, choose $\varphi^{(n+1)} = \varphi^{(n)} + \kappa_n (\varphi^* - \varphi^{(n)})$

WO 2004/104856

PCT/AU2004/000696

- 77 -

where φ^* satisfies $\frac{\partial L(\underline{y} | P_n \gamma^*, \varphi)}{\partial \varphi} = 0$ and κ_n is a damping factor

such that $0 < \kappa_n < 1$.

6. Check convergence. If $\|\gamma^* - \gamma^{(n)}\| < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step 2 above.

5

In another embodiment, step (D11) in the maximisation step may be estimated by replacing $\frac{\partial^2 L}{\partial^2 \gamma_r}$ with its expectation

$$E\left\{\frac{\partial^2 L}{\partial^2 \gamma_r}\right\}.$$

- 10 In one embodiment, the EM algorithm is applied to maximise the posterior distribution when the model is Cox's proportional hazard's model.

To aid in the exposition of the application of the EM
15 algorithm when the model is Cox's proportional hazards model, it is preferred to define "dynamic weights" and matrices based on these weights. The weights are -

$$w_{i,l} = \frac{\zeta_{i,l} \exp(Z_l \beta)}{\sum_{j=1}^N \zeta_{i,j} \exp(Z_j \beta)},$$

$$w_l^* = \sum_{i=1}^N d_i w_{i,l},$$

$$\tilde{w}_l = d_l - w_l^*.$$

20

Matrices based on these weights are -

$$W_i = \begin{pmatrix} w_{i,1} \\ w_{i,2} \\ \vdots \\ w_{i,N} \end{pmatrix},$$

$$\tilde{W} = \begin{pmatrix} \tilde{w}_1 \\ \tilde{w}_2 \\ \vdots \\ \tilde{w}_N \end{pmatrix},$$

$$\Delta(W^*) = \begin{pmatrix} w_1^* & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_N^* \end{pmatrix},$$

$$W^{**} = \sum_{i=1}^N d_i W_i W_i^T$$

5

$$\left. \begin{aligned} \frac{\partial L}{\partial \underline{\beta}} &= Z^T \bar{W} \\ \frac{\partial^2 L}{\partial \underline{\beta}^2} &= Z^T \left(W^{**} - \Delta(W^*) \right) Z = Z^T K Z \end{aligned} \right\} \quad (D12)$$

10

$$\left. \begin{aligned} \frac{\partial L}{\partial \gamma_r} &= P_n^T \frac{\partial L}{\partial \beta_r} = P_n^T Z^T \tilde{W} \\ \frac{\partial^2 L}{\partial \gamma_r^2} &= P_n^T \frac{\partial^2 L}{\partial \beta_r^2} P_n = P_n^T Z^T (W^{**} - \Delta(W^*)) Z P_n = P_n^T Z^T K Z P_n \end{aligned} \right\} \quad (D13)$$

WO 2004/104856

PCT/AU2004/000696

- 79 -

Preferably, when the model is Cox's proportional hazards model the E step and M step of the EM algorithm are as follows:

1. Choose the hyperprior and its parameters b and k . Set
 5 $n=0$, $S_0 = \{1, 2, \dots, p\}$. Let v be the vector with components

$$v_i = \begin{cases} 1-\epsilon, & \text{if } c_i=1 \\ \epsilon, & \text{if } c_i=0 \end{cases}$$

- 10 for some small ϵ , say .001. Define f to be $\log(v/t)$.

If $p \leq N$ compute initial values $\underline{\beta}^*$ by

$$\underline{\beta}^* = (Z^T Z + \lambda I)^{-1} Z^T f$$

- 15 If $p > N$ compute initial values $\underline{\beta}^*$ by

$$\underline{\beta}^* = \frac{1}{\lambda} (I - Z^T (Z Z^T + \lambda I)^{-1} Z) Z^T f$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$.

- 20 2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

Let P_n be a matrix of zeroes and ones such that the nonzero elements $\gamma^{(n)}$ of $\beta^{(n)}$ satisfy

25
$$\begin{aligned} \gamma^{(n)} &= P_n^T \beta^{(n)}, & \beta^{(n)} &= P_n \gamma^{(n)} \\ \gamma &= P_n^T \beta, & \beta &= P_n \gamma \end{aligned}$$

3. Perform the E step by calculating

$$\begin{aligned} Q(\underline{\beta} | \underline{\beta}^{(n)}) &= E \left\{ \log \left(P(\underline{\beta}, \underline{\varphi}, \underline{\tau} | \underline{t}) \right) | \underline{t}, \underline{\beta}^{(n)} \right\} \\ &= L(\underline{t} | \underline{\beta}) - \frac{1}{2} \sum_{i=1}^N \left(\frac{\beta_i}{d_i(\beta^{(n)})} \right)^2 \end{aligned}$$

where L is the log likelihood function of \underline{t} given by Equation (8). Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ we have

$$Q(\underline{y}|\underline{y}^{(n)}) = L(\underline{t} | P_n \underline{y}) - \frac{1}{2} \underline{y}^T \Delta(d(\underline{y}^{(n)})) \underline{y}$$

4. Do the M step. This can be done with Newton Raphson iterations as follows. Set $\gamma_0 = \gamma^{(r)}$ and for $r=0,1,2,\dots$ $\gamma_{r+1} = \gamma_r + \alpha_r \delta_r$ where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1} | \gamma^{(n)}, \phi^{(n)}) > Q(\gamma_r | \gamma^{(n)}, \phi^{(n)})$.

10 For $p \leq N$ use

$$\underline{\delta}_r = \Delta \left(d(\underline{\gamma}^{(n)}) \right) \left(Y^T K Y + I \right)^{-1} \left(Y^T \tilde{W} - \Delta \left(1/d(\underline{\gamma}^{(n)}) \right) \underline{\gamma} \right),$$

where $Y = ZP_n \Delta \left(d(\underline{\gamma}^{(n)}) \right)$.

For $p > N$ use

$$\underline{\delta}_r = \Delta \left(d(\underline{\gamma}^{(n)}) \right) \left(I - Y^T (Y Y^T + K^{-1})^{-1} Y \right) \left(Y^T \bar{W} - \Delta \left(1/d(\underline{\gamma}^{(n)}) \right) \underline{\gamma} \right)$$

Let γ^* be the value of γ_r when some convergence
15 criterion is satisfied e.g. $||\gamma_r - \gamma_{r+1}|| < \epsilon$ (for example 10^{-5}).

5. Define $\beta_{\underline{n}}^* = P_n \gamma^*$, $S_n = \left\{ i : |\beta_i| > \varepsilon_1 \max_j |\beta_j| \right\}$ where ε_1 is a

small constant, say 10^{-5} . This step eliminates variables with very small coefficients.

20 6. Check convergence. If $\|\gamma^* - \gamma^{(n)}\| < \varepsilon_2$ where ε_2 is suitably small then stop, else set $n=n+1$, go to step 2 above and repeat procedure until convergence occurs.

- 81 -

In another embodiment the EM algorithm is applied to maximise the posterior distribution when the model is a parametric survival model.

- 5 In applying the EM algorithm to the parametric survival model, a consequence of equation (11) is that the c_i 's may be treated as Poisson variates with means μ_i and that the last term in equation (11) does not depend on β (although it depends on φ). Note that $\log(\mu_i) = \log(\Lambda(y_i; \varphi)) + X_i \beta$ and so
 10 it is possible to couch the problem in terms of a log-linear model for the Poisson-like mean. Preferably, an iterative maximization of the log-likelihood function is performed where given initial estimates of φ the estimates of β are obtained. Then given these estimates of β , updated
 15 estimates of φ are obtained. The procedure is continued until convergence occurs.

Applying the posterior distribution described above, we note that (for fixed φ)

$$\frac{\partial \log(\mu)}{\partial \beta} = \frac{1}{\mu} \frac{\partial \mu}{\partial \beta} \Leftrightarrow \frac{\partial \mu}{\partial \beta} = \mu \frac{\partial \log(\mu)}{\partial \beta} \text{ and } \frac{\partial \log(\mu_i)}{\partial \beta} = X_i \quad (\text{D14})$$

- 20 Consequently from Equations (11) and (12) it follows that

$$\frac{\partial L}{\partial \beta} = X^T (\underline{c} - \underline{\mu}) \text{ and } \frac{\partial^2 L}{\partial \beta^2} = -X^T \Delta(\underline{\mu}) X.$$

The versions of Equation (12) relevant to the parametric survival models are

$$\left. \begin{aligned} \frac{\partial L}{\partial \gamma_r} &= P_n^T \frac{\partial L}{\partial \beta_r} = P_n^T X^T (\underline{c} - \underline{\mu}) \\ \frac{\partial^2 L}{\partial \gamma_r^2} &= P_n^T \frac{\partial^2 L}{\partial \beta_r^2} P_n = -P_n^T X^T \Delta(\underline{\mu}) X P_n \end{aligned} \right\} \quad (\text{D15})$$

To solve for φ after each M step of the EM algorithm (see

step 5 below) preferably put $\varphi^{(n+1)} = \varphi^{(n+1)} + \kappa_n (\varphi^* - \varphi^{(n)})$ where φ^*

- 82 -

satisfies $\frac{\partial L}{\partial \varphi} = 0$ for $0 < \kappa_n \leq 1$ and β is fixed at the value

obtained from the previous M step.

It is possible to provide an EM algorithm for parameter selection in the context of parametric survival models and microarray data. Preferably, the EM algorithm is as follows:

1. Choose a hyper prior and its parameters b and k eg $b=1e7$ and $k=0.5$. Set $n=0$, $S_0 = \{1, 2, \dots, p\}$ $\varphi^{(initial)} = \varphi^{(0)}$. Let v be the vector with components

$$v_i = \begin{cases} 1-\epsilon, & \text{if } c_i=1 \\ \epsilon, & \text{if } c_i=0 \end{cases}$$

for some small ϵ , say for example .001. Define f to be $\log(v/\Lambda(y, \varphi))$.

If $p \leq N$ compute initial values β^* by $\beta^* = (X^T X + \lambda I)^{-1} X^T f$.
If $p > N$ compute initial values β^* by

$$\beta^* = \frac{1}{\lambda} (I - X^T (X X^T + \lambda I)^{-1} X) X^T f$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$.

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

Let P_n be a matrix of zeroes and ones such that the nonzero elements $\gamma^{(n)}$ of $\beta^{(n)}$ satisfy

$$\begin{aligned} \gamma^{(n)} &= P_n^T \beta^{(n)}, & \beta^{(n)} &= P_n \gamma^{(n)} \\ \gamma &= P_n^T \beta, & \beta &= P_n \gamma \end{aligned}$$

3. Perform the E step by calculating

$$\begin{aligned} Q(\underline{\beta} | \underline{\beta}^{(n)}, \underline{\varphi}^{(n)}) &= E \left\{ \log \left(P(\underline{\beta}, \underline{\varphi}, \tau | \underline{y}) \right) | \underline{y}, \underline{\beta}^{(n)}, \underline{\varphi}^{(n)} \right\} \\ &= L(\underline{y} | \underline{\beta}, \underline{\varphi}^{(n)}) - \frac{1}{2} \sum_{i=1}^N \left(\frac{\beta_i}{d(\beta^{(n)})} \right)^2 \end{aligned}$$

5

where L is the log likelihood function of \underline{y} and $\underline{\varphi}^{(n)}$.

Using $\underline{\beta} = P_n \underline{\gamma}$ and $\underline{\beta}^{(n)} = P_n \underline{\gamma}^{(n)}$ we have

$$Q(\underline{\gamma} | \underline{\gamma}^{(n)}, \underline{\varphi}^{(n)}) = L(\underline{y} | P_n \underline{\gamma}, \underline{\varphi}^{(n)}) - \frac{1}{2} \underline{\gamma}^T \Delta(d(\underline{\gamma}^{(n)})) \underline{\gamma}$$

4. Do the M step. This can be done with Newton Raphson

10 iterations as follows. Set $\underline{\gamma}_0 = \underline{\gamma}^{(r)}$ and for $r=0, 1, 2, \dots$

$\underline{\gamma}_{r+1} = \underline{\gamma}_r + \alpha_r \underline{\delta}_r$ where α_r is chosen by a line search

algorithm to ensure $Q(\underline{\gamma}_{r+1} | \underline{\gamma}^{(n)}, \underline{\varphi}^{(n)}) > Q(\underline{\gamma}_r | \underline{\gamma}^{(n)}, \underline{\varphi}^{(n)})$.

For $p \leq N$ use

$$\underline{\delta}_r = -\Delta(d(\underline{\gamma}^{(n)})) [Y_n^T \Delta(\underline{\mu}) Y_n + I]^{-1} (Y_n^T (\underline{y} - \underline{\mu}) - \Delta(1/d(\underline{\gamma}^{(n)})) \underline{\gamma})$$

where $Y = X P_n \Delta(d(\underline{\gamma}^{(n)}))$.

15 For $p > N$ use

$$\underline{\delta}_r = -\Delta(d(\underline{\gamma}^{(n)})) \left(I - Y^T (Y Y^T + \Delta(1/\underline{\mu}))^{-1} Y \right) \left(Y_n^T (\underline{y} - \underline{\mu}) - \Delta(1/d(\underline{\gamma}^{(n)})) \underline{\gamma} \right)$$

Let $\underline{\gamma}^*$ be the value of $\underline{\gamma}_r$ when some convergence criterion is satisfied e.g. $|| \underline{\gamma}_r - \underline{\gamma}_{r+1} || < \varepsilon$ (for example 10^{-5}).

20 5. Define $\underline{\beta}^* = P_n \underline{\gamma}^*$, $S_n = \left\{ i : |\beta_i| > \varepsilon_1 \max_j |\beta_j| \right\}$ where ε_1 is a small

constant, say 10^{-5} . Set $n=n+1$, choose $\underline{\varphi}^{(n+1)} = \underline{\varphi}^{(n)} + \kappa_n (\underline{\varphi}^* - \underline{\varphi}^{(n)})$

WO 2004/104856

PCT/AU2004/000696

- 84 -

where $\underline{\varphi}^*$ satisfies $\frac{\partial L(\underline{y} | P_n \underline{\gamma}^*, \underline{\varphi})}{\partial \underline{\varphi}} = 0$ and κ_n is a damping

factor such that $0 < \kappa_n < 1$.

6. Check convergence. If $\|\underline{\gamma}^* - \underline{\gamma}^{(n)}\| < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step 2.

In another embodiment, survival times are described by a Weibull survival density function. For the Weibull case $\underline{\varphi}$ is preferably one dimensional and

10

$$\begin{aligned} \Lambda(y; \underline{\varphi}) &= y^\alpha, \\ \lambda(y; \underline{\varphi}) &= \alpha y^{\alpha-1}, \\ \underline{\varphi} &= \alpha \end{aligned}$$

Preferably, $\frac{\partial L}{\partial \alpha} = \frac{N}{\alpha} + \sum_{i=1}^N (c_i - \mu_i) \log(y_i) = 0$ is solved

after each M step so as to provide an updated value of α .

- 15 Following the steps applied for Cox's proportional hazards model, one may estimate α and select a parsimonious subset of parameters from $\underline{\beta}$ that can provide an adequate explanation for the survival times if the survival times follow a Weibull distribution. A numerical example is now given.

20

The preferred embodiment of the present invention will now be described by way of reference only to the following non-limiting example. It should be understood, however, that

25 the example is illustrative only, should not be taken in any way as a restriction on the generality of the invention described above.

Example:

- ```

20 sigma squared 0.6004567
 EM Iteration: 1 expected post: -63.91024 basis fns 41

 sigma squared 0.6037467
 EM Iteration: 2 expected post: -52.76575 basis fns 41
25 sigma squared 0.6081233
 EM Iteration: 3 expected post: -53.10084 basis fns 30

 sigma squared 0.6118665
 EM Iteration: 4 expected post: -53.55141 basis fns 22
30 sigma squared 0.6143482
 EM Iteration: 5 expected post: -53.79887 basis fns 18

 sigma squared 0.6155
 EM Iteration: 6 expected post: -53.91096 basis fns 18

```

- 86 -

sigma squared 0.6159484

EM Iteration: 7 expected post: -53.94735 basis fns 16

sigma squared 0.6160951

5 EM Iteration: 8 expected post: -53.92469 basis fns 14

sigma squared 0.615873

**EM Iteration: 9 expected post: -53.83668 basis fns 13**

```
10 sigma squared 0.6156233
```

EM Iteration: 10 expected post: -53.71836 basis fns 13

sigma squared 0.6156616

EM Iteration: 11 expected post: -53.61035 basis fns 12

15

sigma squared 0.6157966

EM Iteration: 12 expected post: -53.52386 basis fns 12

sigma squared 0.6159524

20 EM Iteration: 13 expected post: -53.47354 basis fns 12

sigma squared 0.6163736

**EM Iteration: 14 expected post: -53.47986 basis fns 12**

```
25 sigma squared 0.6171314
```

**EM Iteration: 15 expected post: -53.53784 basis fns 11**

sigma squared 0.6182353

EM Iteration: 16 expected post: -53.63423 basis fns 11

30

sigma squared 0.6196385

EM Iteration: 17 expected post: -53.75112 basis fns 11

sigma squared 0.621111

```
35 EM Iteration: 18 expected post: -53.86309 basis fns 11
```

sigma squared 0.6224584

For the reduced data set with 201 observations and 10 variables,  $k=0$  and  $b=1e7$

Gives the correct basis functions, namely 1 2 3 4. With  $k=0.25$  and  $b=1e7$ , 7 basis functions were chosen, namely 1 2 3 4 6 8 9. A record of the iterations is given below.

Note that this set also includes the correct set.

misclassification table

```
Class 1 Number of basis functions in model : 6431

```

WO 2004/104856

PCT/AU2004/000696

- 90 -

Iteration 3 : 4 cycles, criterion -9.007843

misclassification matrix

fhat

```
5 f 1 2
 1 23 0
 2 0 22
```

row =true class

10 Class 1 Number of basis functions in model : 508

\*\*\*\*\*

Iteration 4 : 5 cycles, criterion -6.47555

misclassification matrix

fhat

```
15 f 1 2
 1 23 0
 2 0 22
```

row =true class

20

Class 1 Number of basis functions in model : 62

\*\*\*\*\*

Iteration 5 : 6 cycles, criterion -4.126996

25 misclassification matrix

fhat

```
f 1 2
 1 23 0
 2 1 21
```

30 row =true class

Class 1 Number of basis functions in model : 20

\*\*\*\*\*

Iteration 6 : 6 cycles, criterion -3.047699

35

misclassification matrix

fhat



WO 2004/104856

PCT/AU2004/000696

- 91 -

f 1 2

1 23 0

2 1 21

row =true class

5

Class 1 Number of basis functions in model : 12

\*\*\*\*\*

Iteration 7 : 5 cycles, criterion -2.610974

10 misclassification matrix

fhat

f 1 2

1 23 0

2 1 21

15 row =true class

Class 1 : Variables left in model

1 2 3 408 846 6614 7191 8077

regression coefficients

20 28.81413 14.27784 7.025863 -1.086501e-06 4.553004e-09 -

16.25844 0.1412991 -0.04101412

\*\*\*\*\*

Iteration 8 : 5 cycles, criterion -2.307441

25

misclassification matrix

fhat

f 1 2

1 23 0

30 2 1 21

row =true class

Class 1 : Variables left in model

1 2 3 6614 7191 8077

35 regression coefficients

32.66699 15.80614 7.86011 -18.53527 0.1808061 -0.006728619

WO 2004/104856

PCT/AU2004/000696

- 92 -

\*\*\*\*\*

Iteration 9 : 5 cycles, criterion -2.028043

misclassification matrix

5 fhat

f 1 2

1 23 0

2 0 22

row =true class

10

Class 1 : Variables left in model

1 2 3 6614 7191 8077

regression coefficients

36.11990 17.21351 8.599812 -20.52450 0.2232955 -0.0001630341

15

\*\*\*\*\*

Iteration 10 : 6 cycles, criterion -1.808861

misclassification matrix

20 fhat

f 1 2

1 23 0

2 0 22

row =true class

25

Class 1 : Variables left in model

1 2 3 6614 7191 8077

regression coefficients

39.29053 18.55341 9.292612 -22.33653 0.260273 -8.696388e-08

30

\*\*\*\*\*

Iteration 11 : 6 cycles, criterion -1.656129

misclassification matrix

35 fhat

f 1 2

1 23 0

WO 2004/104856

PCT/AU2004/000696

- 93 -

```

 2 0 22
row =true class

Class 1 : Variables left in model
5 1 2 3 6614 7191
 regression coefficients
 42.01569 19.73626 9.90312 -23.89147 0.2882204

10 Iteration 12 : 6 cycles, criterion -1.554494

misclassification matrix
 fhat
f 1 2
15 1 23 0
 2 0 22
row =true class

Class 1 : Variables left in model
20 1 2 3 6614 7191
 regression coefficients
 44.19405 20.69926 10.40117 -25.1328 0.3077712

 Iteration 13 : 6 cycles, criterion -1.487778
25

misclassification matrix
 fhat
f 1 2
 1 23 0
30 2 0 22
row =true class

Class 1 : Variables left in model
 1 2 3 6614 7191
35 regression coefficients
 45.84032 21.43537 10.78268 -26.07003 0.3209974

```

WO 2004/104856

PCT/AU2004/000696

- 94 -

\*\*\*\*\*

Iteration 14 : 6 cycles, criterion -1.443949

misclassification matrix

5 fhat

f 1 2  
1 23 0  
2 0 22

row =true class

10

Class 1 : Variables left in model

1 2 3 6614 7191

regression coefficients

47.03702 21.97416 11.06231 -26.75088 0.3298526

15

\*\*\*\*\*

Iteration 15 : 6 cycles, criterion -1.415

misclassification matrix

20 fhat

f 1 2  
1 23 0  
2 0 22

row =true class

25

Class 1 : Variables left in model

1 2 3 6614 7191

regression coefficients

47.88472 22.35743 11.26136 -27.23297 0.3357765

30

\*\*\*\*\*

Iteration 16 : 6 cycles, criterion -1.395770

misclassification matrix

35 fhat

f 1 2  
1 23 0

```

2 0 22
row =true class

Class 1 : Variables left in model
5 1 2 3 6614 7191
regression coefficients
48.47516 22.62508 11.40040 -27.56866 0.3397475

10 Iteration 17 : 5 cycles, criterion -1.382936

misclassification matrix
fhat
f 1 2
15 1 23 0
2 0 22
row =true class

Class 1 : Variables left in model
20 1 2 3 6614 7191
regression coefficients
48.88196 22.80978 11.49636 -27.79991 0.3424153

25 Iteration 18 : 5 cycles, criterion -1.374340

misclassification matrix
fhat
f 1 2
30 1 23 0
2 0 22
row =true class

Class 1 : Variables left in model
35 1 2 3 6614 7191
regression coefficients
49.16029 22.93629 11.56209 -27.95811 0.3442109

```

WO 2004/104856

PCT/AU2004/000696

- 96 -

\*\*\*\*\*

Iteration 19 : 5 cycles, criterion -1.368567

5 misclassification matrix

fhat

f 1 2

1 23 0

2 0 22

10 row =true class

Class 1 : Variables left in model

1 2 3 6614 7191

regression coefficients

15 49.34987 23.02251 11.60689 -28.06586 0.3454208

\*\*\*\*\*

Iteration 20 : 5 cycles, criterion -1.364684

20 misclassification matrix

fhat

f 1 2

1 23 0

2 0 22

25 row =true class

Class 1 : Variables left in model

1 2 3 6614 7191

regression coefficients

30 49.47861 23.08109 11.63732 -28.13903 0.3462368

\*\*\*\*\*

Iteration 21 : 5 cycles, criterion -1.362068

35 misclassification matrix

fhat

f 1 2

regression coefficients

WO 2004/104856

PCT/AU2004/000696

- 98 -

49.6649 23.16588 11.68137 -28.2449 0.3474096

\*\*\*\*\*

Iteration 24 : 4 cycles, criterion -1.358314

5

misclassification matrix

fhat

f 1 2

1 23 0

10 2 0 22

row =true class

Class 1 : Variables left in model

1 2 3 6614 7191

15 regression coefficients

49.69192 23.17818 11.68776 -28.26025 0.3475789

\*\*\*\*\*

Iteration 25 : 4 cycles, criterion -1.357772

20

misclassification matrix

fhat

f 1 2

1 23 0

25 2 0 22

row =true class

Class 1 : Variables left in model

1 2 3 6614 7191

30 regression coefficients

49.71017 23.18649 11.69208 -28.27062 0.3476932

\*\*\*\*\*

Iteration 26 : 4 cycles, criterion -1.357407

35

misclassification matrix

fhat



- 99 -

```
f 1 2
 1 23 0
 2 0 22
row =true class

5
Class 1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.72251 23.19211 11.695 -28.27763 0.3477704

10

Iteration 27 : 4 cycles, criterion -1.35716

misclassification matrix

15 fhat
f 1 2
 1 23 0
 2 0 22
row =true class

20
Class 1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.73084 23.19590 11.69697 -28.28237 0.3478225

25

Iteration 28 : 3 cycles, criterion -1.356993

misclassification matrix

30 fhat
f 1 2
 1 23 0
 2 0 22
row =true class

35
Class 1 : Variables left in model
1 2 3 6614 7191
```

**PCT/AU2004/000696**

- 100 -

regression coefficients

49.73646 23.19846 11.6983 -28.28556 0.3478577

\*\*\*\*\*

5 Iteration 29 : 3 cycles, criterion -1.356881

misclassification matrix

that

| <b>f</b> | <b>1</b> | <b>2</b> |
|----------|----------|----------|
|----------|----------|----------|

10            1 23    0

2 0 22

```
row =true class
```

**Class 1 : Variables left in model**

15 1 2 3 6614 7191

regression coefficients

49.74026 23.20019 11.6992 -28.28772 0.3478814

\*\*\*\*\*

```
20 Iteration 30 : 3 cycles, criterion -1.356805
```

**misclassification matrix**

that

**f 1 2**

25            1 23    0

**2 0 22**

```
row =true class
```

**Class 1 : Variables left in model**

30 1 2 3 6614 7191

regression coefficients

49.74283 23.20136 11.69981 -28.28918 0.3478975

1

35

misclassification table

**pred**

- 101 -

```

y 1 2 3 4
 1 4 0 0 0
 2 0 3 0 0
 3 0 0 4 0
5 4 0 0 0 4

```

Identifiers of variables left in ordered categories model  
6611 7188

- 10 Ordered categories example  
Luo prostate data 15 samples 50 genes. For  $k=0$  and  $b=1e7$  we  
get the following results

```

misclassification table
15 pred
 y 1 2 3 4
 1 4 0 0 0
 2 0 2 1 0
 3 0 0 4 0
20 4 0 0 0 4

```

Identifiers of variables left in ordered categories model  
1

- 25 For  $k=0.25$  and  $b=1e7$  we get the following results

```

misclassification table
 pred
30 y 1 2 3 4
 1 4 0 0 0
 2 0 3 0 0
 3 0 0 4 0
 4 0 0 0 4

```

- 35 Identifiers of variables left in ordered categories model  
1 42

WO 2004/104856

PCT/AU2004/000696

- 102 -

A record of the iterations for  $k=0.25$  and  $b=1e7$  is given below

\*\*\*\*\*

5 Iteration 1 : 19 cycles, criterion -0.4708706

misclassification matrix

fhat

f 1 2  
10 1 23 0  
2 0 22

row =true class

Class 1 Number of basis functions in model : 53

15 \*\*\*\*\*

Iteration 2 : 7 cycles, criterion -1.536822

misclassification matrix

fhat

20 f 1 2  
1 23 0  
2 0 22

row =true class

25 Class 1 Number of basis functions in model : 53

\*\*\*\*\*

Iteration 3 : 5 cycles, criterion -2.032919

misclassification matrix

fhat

30 f 1 2  
1 23 0  
2 0 22

row =true class

35

Class 1 Number of basis functions in model : 42

\*\*\*\*\*

- 103 -

Iteration 4 : 5 cycles, criterion -2.132546

misclassification matrix

fhat

5 f 1 2

1 23 0

2 0 22

row =true class

10 Class 1 Number of basis functions in model : 18

\*\*\*\*\*

Iteration 5 : 5 cycles, criterion -1.978462

misclassification matrix

fhat

15 f 1 2

1 23 0

2 0 22

row =true class

20

Class 1 Number of basis functions in model : 13

\*\*\*\*\*

Iteration 6 : 5 cycles, criterion -1.668212

25 misclassification matrix

fhat

f 1 2

1 23 0

2 0 22

30 row =true class

Class 1 : Variables left in model

1 2 3 4 10 41 43 45

regression coefficients

35 34.13253 22.30781 13.04342 -16.23506 0.003213167 0.006582334

-0.0005943874 -3.557023

WO 2004/104856

PCT/AU2004/000696

- 104 -

\*\*\*\*\*

Iteration 7 : 5 cycles, criterion -1.407871

misclassification matrix

5 fhat

f 1 2

1 23 0

2 0 22

row =true class

10

Class 1 : Variables left in model

1 2 3 4 10 41 43 45

regression coefficients

36.90726 24.69518 14.61792 -17.16723 1.112172e-05 5.949931e-

15 06 -3.892181e-08 -4.2906

\*\*\*\*\*

Iteration 8 : 5 cycles, criterion -1.244166

20 misclassification matrix

fhat

f 1 2

1 23 0

2 0 22

25 row =true class

Class 1 : Variables left in model

1 2 3 4 10 45

regression coefficients

30 39.15038 26.51011 15.78594 -17.99800 1.125451e-10 -4.799167

\*\*\*\*\*

Iteration 9 : 5 cycles, criterion -1.147754

35 misclassification matrix

fhat

f 1 2

```

1 23 0
2 0 22
row =true class

5 Class 1 : Variables left in model
1 2 3 4 45
regression coefficients
40.72797 27.73318 16.56101 -18.61816 -5.115492

10 *****
Iteration 10 : 5 cycles, criterion -1.09211

misclassification matrix
fhat
15 f 1 2
1 23 0
2 0 22
row =true class

20 Class 1 : Variables left in model
1 2 3 4 45
regression coefficients
41.74539 28.49967 17.04204 -19.03293 -5.302421

25 *****
Iteration 11 : 5 cycles, criterion -1.060238

misclassification matrix
fhat
30 f 1 2
1 23 0
2 0 22
row =true class

35 Class 1 : Variables left in model
1 2 3 4 45
regression coefficients

```

WO 2004/104856

PCT/AU2004/000696

- 106 -

42.36866 28.96076 17.32967 -19.29261 -5.410496

\*\*\*\*\*

Iteration 12 : 5 cycles, criterion -1.042037

5

misclassification matrix

fhat

f 1 2

1 23 0

10

2 0 22

row =true class

Class 1 : Variables left in model

1 2 3 4 45

15

regression coefficients

42.73908 29.23176 17.49811 -19.44894 -5.472426

\*\*\*\*\*

Iteration 13 : 5 cycles, criterion -1.031656

20

misclassification matrix

fhat

f 1 2

1 23 0

25

2 0 22

row =true class

Class 1 : Variables left in model

1 2 3 4 45

30

regression coefficients

42.95536 29.38894 17.59560 -19.54090 -5.507787

\*\*\*\*\*

Iteration 14 : 4 cycles, criterion -1.025738

35

misclassification matrix

fhat



- 107 -

```
f 1 2
 1 23 0
 2 0 22
row =true class

5
Class 1 : Variables left in model
 1 2 3 4 45
regression coefficients
43.08034 29.47941 17.65163 -19.59428 -5.527948

10

Iteration 15 : 4 cycles, criterion -1.022366

misclassification matrix

15 fhat
 f 1 2
 1 23 0
 2 0 22
 row =true class

20
Class 1 : Variables left in model
 1 2 3 4 45
regression coefficients
43.15213 29.53125 17.68372 -19.62502 -5.539438

25

Iteration 16 : 4 cycles, criterion -1.020444

misclassification matrix

30 fhat
 f 1 2
 1 23 0
 2 0 22
 row =true class

35
Class 1 : Variables left in model
 1 2 3 4 45
```

WO 2004/104856

PCT/AU2004/000696

- 108 -

regression coefficients

43.19322 29.56089 17.70206 -19.64265 -5.545984

\*\*\*\*\*

5 Iteration 17 : 4 cycles, criterion -1.019349

misclassification matrix

fhat

f 1 2

10 1 23 0

2 0 22

row =true class

Class 1 : Variables left in model

15 1 2 3 4 45

regression coefficients

43.21670 29.57780 17.71252 -19.65272 -5.549713

\*\*\*\*\*

20 Iteration 18 : 3 cycles, criterion -1.018725

misclassification matrix

fhat

f 1 2

25 1 23 0

2 0 22

row =true class

Class 1 : Variables left in model

30 1 2 3 4 45

regression coefficients

43.23008 29.58745 17.71848 -19.65847 -5.551837

\*\*\*\*\*

35 Iteration 19 : 3 cycles, criterion -1.01837

misclassification matrix

- 109 -

```
 fhat
f 1 2
 1 23 0
 2 0 22
5 row =true class

Class 1 : Variables left in model
1 2 3 4 45
regression coefficients
10 43.23772 29.59295 17.72188 -19.66176 -5.553047

Iteration 20 : 3 cycles, criterion -1.018167

15 misclassification matrix
 fhat
f 1 2
 1 23 0
 2 0 22
20 row =true class

Class 1 : Variables left in model
1 2 3 4 45
regression coefficients
25 43.24208 29.59608 17.72382 -19.66363 -5.553737

Iteration 21 : 3 cycles, criterion -1.018052

30 misclassification matrix
 fhat
f 1 2
 1 23 0
 2 0 22
35 row =true class

Class 1 : Variables left in model
```

- 110 -

```
1 2 3 4 45
regression coefficients
43.24456 29.59787 17.72493 -19.66469 -5.55413

5 *****
Iteration 22 : 3 cycles, criterion -1.017986

misclassification matrix
 fhat
10 f 1 2
 1 23 0
 2 0 22
row =true class

15 Class 1 : Variables left in model
1 2 3 4 45
regression coefficients
43.24598 29.59889 17.72556 -19.6653 -5.554354

20 1

misclassification table
 pred
25 y 1 2 3 4
 1 4 0 0 0
 2 0 3 0 0
 3 0 0 4 0
 4 0 0 0 4
Identifiers of variables left in ordered categories model
30 1 42
```

- 111 -

THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:

1. A method of identifying a subset of components of a system based on data obtained from the system using at least one training sample from the system, the method comprising the steps of:
  - obtaining a linear combination of components of the system and weightings of the linear combination of components, the weightings having values based on data obtained from the at least one training sample, the at least one training sample having a known feature;
  - obtaining a model of a probability distribution of the known feature, wherein the model is conditional on the linear combination of components;
  - obtaining a prior distribution for the weighting of the linear combination of the components, the prior distribution comprising a hyperprior having a high probability density close to zero, the hyperprior being such that it is not a Jeffreys hyperprior;
  - combining the prior distribution and the model to generate a posterior distribution; and
  - identifying the subset of components based on a set of the weightings that maximise the posterior distribution.
2. The method as claimed in claim 1, wherein the step of obtaining the linear combination comprises the step of using a Bayesian statistical method to estimate the weightings.
3. The method as claimed in claim 1 or 2, further comprising the step of making an apriori assumption that a majority of the components are unlikely to be components that will form part of the subset of components.
4. The method as claimed in any one of the preceding claims, wherein the hyperprior comprises one or more adjustable parameters that enable the prior distribution near zero to be varied.

5. The method as claimed in any one of the preceding claims, wherein the model comprise a mathematical equation in the form of a likelihood function that provides the probability distribution based on data obtained from the at least one training sample.

6. The method as claimed in claim 5, wherein the likelihood function is based on a previously described model  
10 for describing some probability distribution.

7. The method as claimed in any one of the preceding claims, wherein the step of obtaining the model comprises the step of selecting the model from a group comprising a multinomial or binomial logistic regression, generalised linear model, Cox's proportional hazards model, accelerated failure model and parametric survival model.

8. The method as claimed in claim 7, wherein the model  
20 based on the multinomial or binomial logistical regression  
is in the form of:

$$L = \prod_{i=1}^n \left[ \prod_{g=1}^{G-1} \left\{ \frac{e^{x_i^T \beta_g}}{\left( 1 + \sum_{g=1}^{G-1} e^{x_i^T \beta_g} \right)} \right\}^{e_{ik}} \left\{ \frac{1}{1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}} \right\}^{e_{iG}} \right]$$

25 9. The method as claimed in claim 7, wherein the model based on the generalised linear model is in the form of:

$$L = \log p(y | \beta, \phi) = \sum_{i=1}^N \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

10. The method as claimed in claim 7, wherein the model  
30 based on the Cox's proportional hazards model is in the form

- 113 -

of:

$$l(\underline{t} | \underline{\beta}) = \prod_{j=1}^N \left( \frac{\exp(Z_j \underline{\beta})}{\sum_{i \in \mathcal{R}_j} \exp(Z_i \underline{\beta})} \right)^{d_j}$$

11. The method as claimed in claim 7, wherein the model based on the Parametric Survival model is in the form of:

5

$$L = \sum_{i=1}^N \left\{ c_i \log(\mu_i) - \mu_i + c_i \left( \log \left( \frac{\lambda(y_i)}{\Lambda(y_i; \varphi)} \right) \right) \right\}$$

12. The method as claimed in any one of the preceding claims, wherein the step of identifying the subset of components comprises the step of using an iterative  
10 procedure such that the probability density of the posterior distribution is maximised.

13. The method as claimed in claim 12, wherein the iterative procedure is an EM algorithm.

15

14. A method for identifying a subset of components of a subject which are capable of classifying the subject into one of a plurality of predefined groups, wherein each group is defined by a response to a test treatment, the method  
20 comprising the steps of:

exposing a plurality of subjects to the test treatment and grouping the subjects into response groups based on responses to the treatment;

measuring components of the subjects; and  
25 identifying a subset of components that is capable of classifying the subjects into response groups using a statistical analysis method.

15. The method as claimed in claim 14, wherein the statistical analysis method comprises the method as claimed in any one of claims 1 to 13.

5 16. An apparatus for identifying a subset of components of  
a subject, the subset being capable of being used to  
classify the subject into one of a plurality of predefined  
response groups wherein each response group, is formed by  
exposing a plurality of subjects to a test treatment and  
10 grouping the subjects into response groups based on the  
response to the treatment, the apparatus comprising:

an input for receiving measured components of the subjects; and

processing means operable to identify a subset of  
15 components that is capable of being used to classify the  
subjects into response groups using a statistical analysis  
method.

17. The apparatus as claimed in claim 16, wherein the  
20 statistical analysis method comprises the method as claimed  
in any one of claims 1 to 15.

18. A method for identifying a subset of components of a subject that is capable of classifying the subject as being responsive or non-responsive to treatment with a test compound, the method comprising the steps of:

exposing a plurality of subjects to the test compound and grouping the subjects into response groups based on each subjects response to the test compound;

30 measuring components of the subjects; and

identifying a subset of components that is capable of being used to classify the subjects into response groups using a statistical analysis method.

35 19. The method as claimed in claim 18, wherein the statistical analysis method comprises the method as claimed in any one of claims 1 to 13.



- 115 -

20. An apparatus for identifying a subset of components of a subject, the subset being capable of being used to classify the subject into one of a plurality of predefined response groups wherein each response group is formed by exposing a plurality of subjects to a compound and grouping the subjects into response groups based on the response to the compound, the apparatus comprising;

an input operable to receive measured components of the subjects;

processing means operable to identify a subset of components that is capable of classifying the subjects into response groups using a statistical analysis method.

21. The apparatus as claimed in claim 20, wherein the statistical analysis method comprises the method as claimed in any one of claims 1 to 15.

22. An apparatus for identifying a subset of components of a system from data generated from the system from a plurality of samples from the system, the subset being capable of being used to predict a feature of a test sample, the apparatus comprising:

a processing means operable to:

obtain a linear combination of components of the system and obtain weightings of the linear combination of components, each of the weightings having a value based on data obtained from at least one training sample, the at least one training sample having a known feature;

obtaining a model of a probability distribution of a second feature, wherein the model is conditional on the linear combination of components;

obtaining a prior distribution for the weightings of the linear combination of the components, the prior distribution comprising an adjustable hyperprior which allows the prior probability mass close to zero to be varied wherein the hyperprior is not a Jeffrey's hyperprior;

- 116 -

combining the prior distribution and the model to generate a posterior distribution; and identifying the subset of components having component weights that maximize the posterior distribution.

5

23. The apparatus as claimed in claim 22, wherein the processing means comprises a computer arranged to execute software.

10 24. A computer program which, when executed by a computing apparatus, allows the computing apparatus to carry out the method as claimed in any one of claims 1 to 13.

15 25. A computer readable medium comprising the computer program as claimed in claim 24.

26. A method of testing a sample from a system to identify a feature of the sample, the method comprising the steps of testing for a subset of components that are diagnostic of the feature, the subset of components having been determined by using the method as claimed in any one of claims 1 to 15.

27. The method as claimed in claim 26, wherein the system is a biological system.

25

28. An apparatus for testing a sample from a system to determine a feature of the sample, the apparatus comprising means for testing for components identified in accordance with the method as claimed in any one of claims 1 to 15.

30

29. A computer program which, when executed by on a computing device, allows the computing device to carry out a method of identifying components from a system that are capable of being used to predict a feature of a test sample from the system, and wherein a linear combination of components and component weights is generated from data generated from a plurality of training samples, each

35

- 117 -

training sample having a known feature, and a posterior distribution is generated by combining a prior distribution for the component weights comprising an adjustable hyperprior which allows the probability mass close to zero to be varied wherein the hyperprior is not a Jeffrey's hyperprior, and a model that is conditional on the linear combination, to estimate component weights which maximise the posterior distribution.

10 30. A method of identifying a subset of components of a biological system, the subset being capable of predicting a feature of a test sample from the biological system, the method comprising the steps of:

15 obtaining a linear combination of components of the system and weightings of the linear combination of components, each of the weightings having a value based on data obtained from at least one training sample, the at least one training sample having a known feature;

20 obtaining a model of a probability distribution of the known feature, wherein the model is conditional on the linear combination of components;

25 obtaining a prior distribution for the weightings of the linear combination of the components, the prior distribution comprising an adjustable hyperprior which allows the probability mass close to zero to be varied;

combining the prior distribution and the model to generate a posterior distribution; and

identifying the subset of components based on the weightings that maximize the posterior distribution.

30

DATED this 26<sup>th</sup> day of May 2004

CSIRO

By their Patent Attorneys

GRIFFITH HACK

35

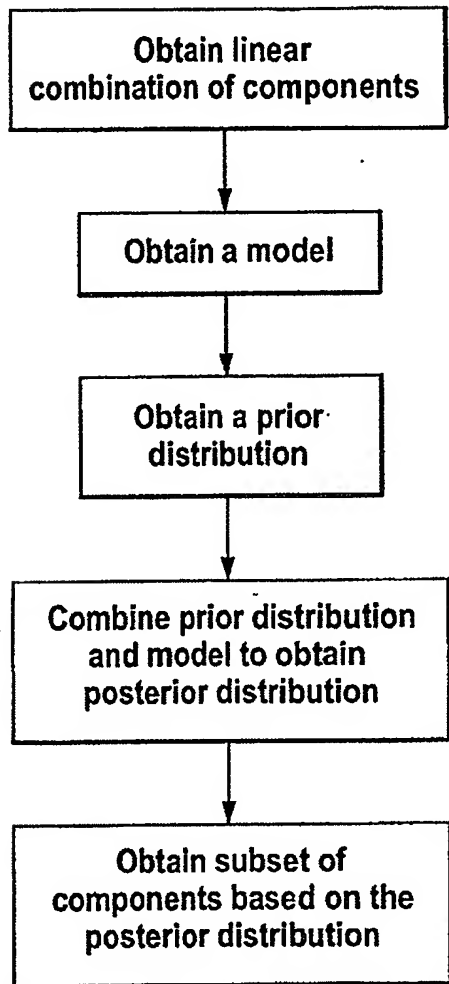


Fig. 1

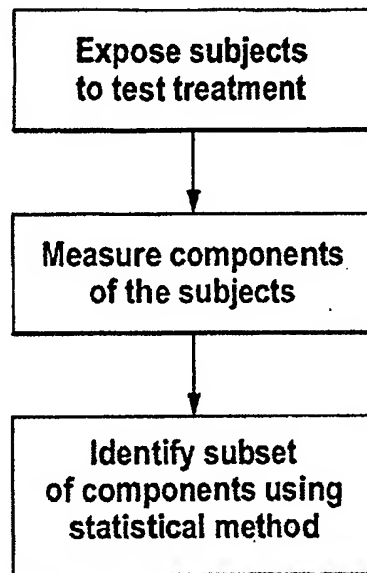


Fig. 2

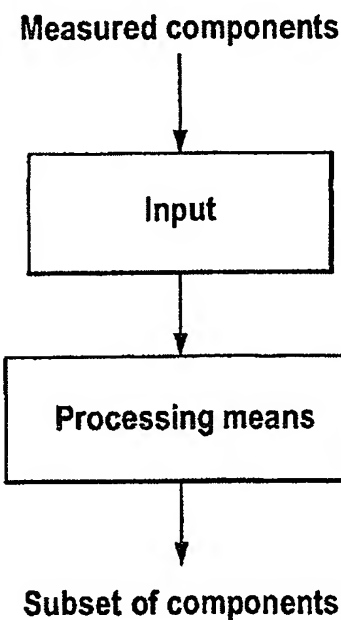


Fig. 3

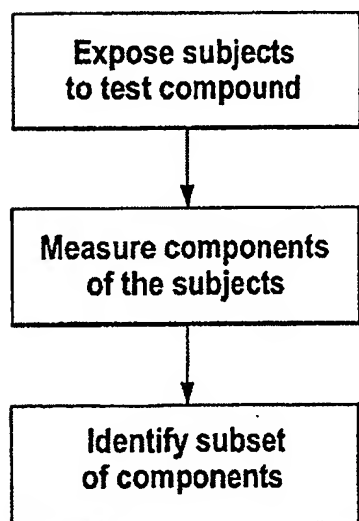


Fig. 4

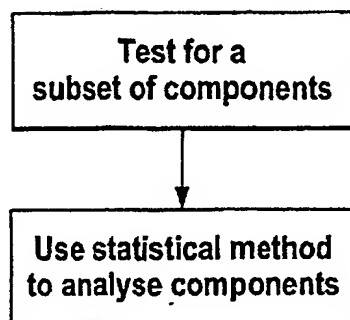


Fig. 5

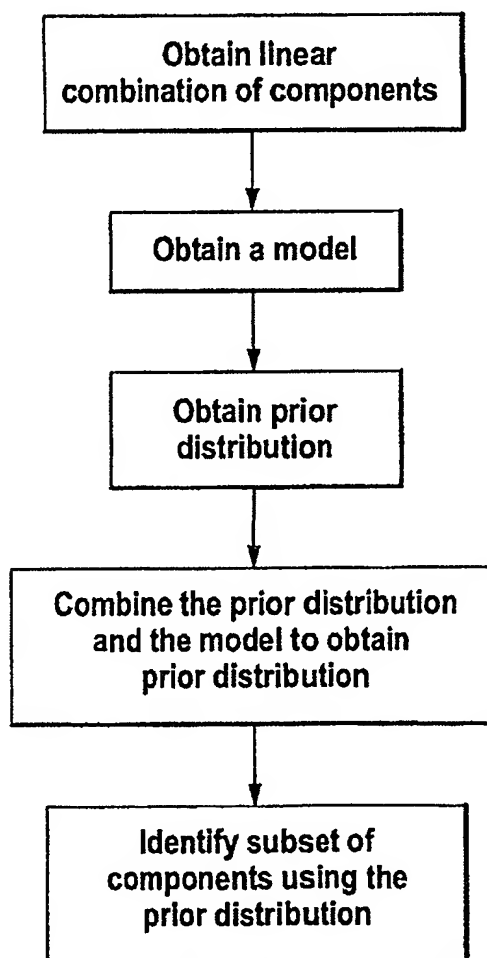


Fig. 6

# INTERNATIONAL SEARCH REPORT

International application No.  
**PCT/AU2004/000696**

## A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. 7: G06F 17/18, G06F 19/00, G06F 159:00

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)  
WPAT, IEEE XPLORE, USPTO (prior, distribution, subset, group, response, posterior+)

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                          | Relevant to claim No.           |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------|
| T         | FIGUEIREDO M.A.T. "Adaptive sparseness for supervised learning" IEEE Transactions on Pattern Analysis and Machine Intelligence September 2003 Sections 3.5, 3.6                                             | 1                               |
| X<br>Y    | WO 2001/018667 A2 (MICROSOFT CORPORATION) 15 March 2001<br>Abstract, page 2 lines 17-22, page 8 line 13, page 11 lines 1-6<br>Whole document                                                                | 1,2,22,23,25,29<br>14,16,18,20, |
| X<br>Y    | WO 2003/034270 A1 (Commonwealth Scientific and Industrial Research Organisation) 24 April 2003, Abstract, Claim 26, page 18 first full paragraph To page 19 "fifth aspect" paragraphs, page 128 lines 29-32 | 14,16,18,20<br>1-30             |
| Y         | WO 2002/087431 A1 (STRUCTURAL BIOINFORMATICS, INC. et al.) 7 November 2002, Abstract                                                                                                                        | 14,16,18,20                     |

☒ Further documents are listed in the continuation of Box C

☒ See patent family annex

\* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search  
20 July 2004

Date of mailing of the international search report

30 JUL 2004

Name and mailing address of the ISA/AU

AUSTRALIAN PATENT OFFICE  
PO BOX 200, WODEN ACT 2606, AUSTRALIA  
E-mail address: pct@ipaustalia.gov.au  
Facsimile No. (02) 6285 3929

Authorized officer

DALE SIVER

Telephone No : (02) 6283 2196

# INTERNATIONAL SEARCH REPORT

International application No.  
PCT/AU2004/000696

| C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT |                                                                                                                                                                |                       |
|-------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Category*                                             | Citation of document, with indication, where appropriate, of the relevant passages                                                                             | Relevant to claim No. |
| Y                                                     | FIGUEIREDO M.A.T. "Bayesian learning of sparse classifiers" IEEE Proceedings of Conference on Computer Vision and Pattern Recognition 2001, Sections 1.6 , 3.2 | 1-30                  |
| Y                                                     | US 6059724 A (CAMPELL et al.) 9 May 2000<br>Abstract, column 25 line 53 to column 26 line3                                                                     | 14,16,18,20           |
| A                                                     | US 64083321 A (PLATT) 18 June 2002<br>Abstract                                                                                                                 |                       |

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/AU2004/000696

This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

| Patent Document Cited in Search Report                                                                                    |          |      |            | Patent Family Member |         |    |         |
|---------------------------------------------------------------------------------------------------------------------------|----------|------|------------|----------------------|---------|----|---------|
| WO                                                                                                                        | 0118667  | AU   | 74736/00   | EP                   | 1226512 | US | 6633857 |
| WO                                                                                                                        | 03034270 | EP   | 1436726    |                      |         |    |         |
| WO                                                                                                                        | 02087431 | US   | 2003065535 |                      |         |    |         |
| US                                                                                                                        | 6059724  | AU   | 61514/98   | BR                   | 9807366 | CA | 2280042 |
|                                                                                                                           |          | EP   | 0973435    | WO                   | 9835609 |    |         |
| US                                                                                                                        | 6408321  | NONE |            |                      |         |    |         |
| Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001. |          |      |            |                      |         |    |         |
| END OF ANNEX                                                                                                              |          |      |            |                      |         |    |         |